STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2747–2764
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2151

A review of methods for futility stopping based on conditional power

John M. Lachin*,†

The Biostatistics Center, Departments of Epidemiology and Biostatistics, and Statistics, The George Washington University, 6110 Executive Boulevard, Suite 750, Rockville, MD 20852, U.S.A.

SUMMARY

Conditional power (CP) is the probability that the final study result will be statistically significant, given the data observed thus far and a specific assumption about the pattern of the data to be observed in the remainder of the study, such as assuming the original design effect, or the effect estimated from the current data, or under the null hypothesis. In many clinical trials, a CP computation at a pre-specified point in the study, such as mid-way, is used as the basis for early termination for futility when there is little evidence of a beneficial effect. Brownian motion can be used to describe the distribution of the interim Z-test value, the corresponding B-value, and the CP values under a specific assumption about the future data. A stopping boundary on the CP value specifies an equivalent boundary on the B-value from which the probability of stopping for futility can then be computed based on the planned study design (sample size and duration) and the assumed true effect size. This yields expressions for the total type I and II error probabilities. As the probability of stopping increases, the probability of a type I error α decreases from the nominal desired level (e.g. 0.05) while the probability of a type II error β increases from the level specified in the study design. Thus a stopping boundary on the B-value can be determined such that the inflation in type II error probability is controlled at a desired level. An iterative procedure is also described that determines a stopping boundary on the B-value and a final test critical Z-value with specified type I and II error probabilities. The implementation in conjunction with a group sequential analysis for effectiveness is also described. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: conditional power; stochastic curtailing; interim analysis; Brownian motion; futility

1. INTRODUCTION

Halperin *et al.* [1] describe 'stochastic curtailing' as a flexible approach to monitor the emerging results in a clinical trial. Stochastic curtailing refers to a decision to terminate the trial

*Correspondence to: John M. Lachin, The Biostatistics Center, Departments of Epidemiology and Biostatistics, and Statistics, The George Washington University, 6110 Executive Boulevard, Suite 750, Rockville, MD 20852, U.S.A.
†E-mail: jml@biostat.bsc.gwu.edu

Contract/grant sponsor: National Institute of Diabetes, Digestive and Kidney Diseases

Received 27 March 2004
Accepted 8 November 2004

based on an assessment of the *conditional power* (CP), defined as the conditional probability that the final result will exceed a critical value given the data observed thus far, and an assumption about the trend of the to be observed in the remainder of the study. Common assumptions include the original design effect, or the effect estimated from the current data, or under the null hypothesis. Lan *et al.* [2] present bounds on the type I and II error probabilities when a study is monitored continuously over time and then stopped early based on high or low, respectively, conditional power levels. Davis and Hardy [3] provide less conservative bounds when the data are evaluated at discrete points in time. Davis and Hardy [4] describe the relationship between decision rules for early termination based on conditional power and other group sequential procedures. In essence, a decision to terminate early and declare significance based on conditional power is equivalent to using the Lan and DeMets [5] group sequential procedure with a specific alpha spending function.

While a conditional power computation could be used as the basis for terminating a trial when a positive effect emerges, a group sequential procedure is usually employed for such decisions. Thus, a conditional power assessment is frequently used to assess the futility of continuing a trial when there is little evidence of a positive effect. In some studies such monitoring for conditional power is done in an *ad hoc* manner, whereas in others a futility stopping criterion is specified. Some studies, particularly in industry, specify that a mid-study assessment be conducted.

Lan and Wittes [6] show that CP levels are readily computed using the B-value, a function of the Z-test value, that follows standard Brownian motion. Herein, the properties of Brownian motion are used to describe the distribution of conditional power values under a specific assumption regarding the trend in the future data. The resulting expressions can be used to assess the probability that a study could be stopped for futility based on the planned study design (sample size) and a specific assumed effect size. Expressions for the type I and II error probabilities are presented for the case where the study plan includes an interim futility analysis at a pre-specified point in time. While it is well known that futility termination will reduce the type I error probability α, and inflate the type II error probability β, it is shown herein that such termination may markedly increase β in direct proportion to the probability of stopping. An iterative procedure is described to determine the final test critical value and the futility stopping boundary such that specified type I and II error probabilities are maintained. Implementation in conjunction with a group sequential procedure for effectiveness is also described.

2. DISTRIBUTION OF INTERIM VALUES

Let Z_t denote the Z-test value at 'information time' t representing the fraction of information accrued in the study at the time of an interim analysis. For a sequence of test statistics with 'independent increments', Lan and Wittes [6] show that conditional power is readily computed using the B-value, $B_t = Z_t\sqrt{t}, 0 < t \leqslant 1$. Lan and Zucker [7] and Lan *et al.* [8] describe the considerations in the quantification of information for many commonly used statistics. Wu and Lan [9], among others, do likewise for various regression models for longitudinal data including the non-linear random effects models. In some cases, such as a test for a difference in means or proportions, the total information is a function of the planned total sample sizes in the experimental and control groups: N_e and N_c, respectively. If n_e and n_c are the

accrued sample sizes at an interim analysis, then the fraction of information at that time is

$$t = \frac{[n_e^{-1} + n_c^{-1}]^{-1}}{[N_e^{-1} + N_c^{-1}]^{-1}} \tag{1}$$

where $t = n/N$ if the two groups are of equal size at the interim and final analyses.

For an analysis of survival data using the logrank test, the total information is a function of the expected total number of events in each group, $E(D_e)$ and $E(D_c)$ [10]. Then for an interim look at which d_e and d_c events have been observed, the fraction of information is

$$t = \frac{[d_e^{-1} + d_c^{-1}]^{-1}}{[E(D_e)^{-1} + E(D_c)^{-1}]^{-1}} \tag{2}$$

Because the total number of events depends on the hazards in each group, or the control group hazard and the relative risk (hazard), this expression is usually evaluated under the null hypothesis cf. References [7, 8]. In this case it is adequate to simply consider the fraction of the expected number of events in the control group $t = d_c/E(D_c)$.

Following Lan and Wittes [6], and also Lan and Zucker [7], the trend in the data is reflected by the *B*-value process over time that is a function of the Brownian motion drift parameter θ. Consider a statistic S that is normally distributed, at least asymptotically, as $S \sim \mathrm{N}(0, \sigma_0^2/N)$ with mean 0 under the null hypothesis (H_0) *versus* $S \sim \mathrm{N}(\phi, \sigma_1^2/N)$ with mean ϕ under the alternative (H_1), where ϕ is the difference in parameters (e.g. means) between the two groups. In some cases, such as the test for proportions, the variance is a function of the expectations in the two groups, in which case $\sigma_0^2 \neq \sigma_1^2$. The null hypothesis would be tested using the test statistic $Z = S\sqrt{N}/\sigma_0$, and the hypothesis rejected at level α against a one or two-sided alternative as appropriate using the critical value $Z_{1-\alpha}$ or $Z_{1-\alpha/2}$, respectively. The equation for the required N is

$$N = \left[\frac{Z_{1-\alpha_D}\sigma_0 + Z_{1-\beta}\sigma_1}{\phi}\right]^2 \tag{3}$$

cf. Reference [11], where α_D is the type I error probability specified under the study design, β is the type II error probability, and $1 - \beta = \Phi(Z_{1-\beta})$ is the power to detect an effect size ϕ. For a two-sided test, $Z_{1-\alpha_D/2}$ is substituted for $Z_{1-\alpha_D}$.

The drift parameter can be described in terms of the non-centrality parameter for the test as

$$\theta = E[Z|H_1] = \frac{\sqrt{N}|\phi|}{\sigma_0} = \frac{Z_{1-\alpha_D}\sigma_0 + Z_{1-\beta}\sigma_1}{\sigma_0}$$

$$\cong Z_{1-\alpha_D} + Z_{1-\beta} \quad \text{when } \sigma_0 \cong \sigma_1 \tag{4}$$

where (3) is substituted for N. The latter equality applies when $\sigma_0 = \sigma_1$, as in the case of mean differences, or applies approximately in other cases under a local alternative. Lachin [11] shows that the difference in variances is negligible in most common applications including the tests for proportions and for exponential survival. Thus a study design that provides 85 per cent power ($Z_{1-\beta} = 1.04$) with a two-sided 0.05 test ($Z_{1-\alpha_D/2} = 1.96$) to detect a specified effect (ϕ) with a given N yields $\theta = 3$ (2.9964 to be precise). Under the null hypothesis of no effect, $1 - \beta = \alpha_D$ and $\theta = 0$.

Lachin [11], among others, presents the expression for the non-centrality parameter and effect size for many widely used test statistics including tests for means, proportions and hazard rates. For the logrank test under a proportional hazards model with hazards $\lambda_c > \lambda_e$ in two groups of equal size, ϕ can be expressed either as

$$\begin{aligned} &\phi = \lambda_c - \lambda_e, \quad \sigma^2 = \frac{2\lambda_c^2}{E(D_c)\,|\,\lambda_c} + \frac{2\lambda_e^2}{E(D_e|\lambda_e)} \quad \text{or as} \\ &\phi = \log(\lambda_c/\lambda_e), \quad \sigma^2 = 2E(D_c|\lambda_c)^{-1} + 2E(D_e|\lambda_e)^{-1} \end{aligned} \tag{5}$$

where $E(D_i|\lambda_i)$ is the expected number of events in each group ($i = c, e$) given the assumed hazard rate and the distribution of exposure (follow-up duration) times based on the planned period and patterns of recruitment, follow-up and losses to follow-up over time. Lachin and Foulkes [12] show that the resulting expressions for sample size or power provide nearly equivalent computations.

The B-value process is expected to follow a trajectory from the point $(0, 0)$ to the point $(1, \theta)$. At information time t, the current estimate of the drift parameter is $\hat{\theta}_t = B_t/t$. Then asymptotically it follows cf. References [6, 7] that

$$\begin{aligned} B_t &= Z_t\sqrt{t} \sim \mathrm{N}(t\theta, t) \\ Z_t &\sim \mathrm{N}(\theta\sqrt{t}, 1) \\ \hat{\theta}_t &= B_t/t \sim \mathrm{N}(\theta, 1/t) \\ U_t &= B_1 - B_t \sim \mathrm{N}[\theta(1-t), 1-t] \end{aligned} \tag{6}$$

where $B_1 = Z_1$ and U_t is the random independent increment in the process yet to be observed over the interval $(t, 1]$. These expressions can be used to describe the distribution of the test statistic Z_t, B-value B_t or the drift parameter estimate $\hat{\theta}_t$ at any interim analysis at time $t \in (0, 1]$ when the true drift parameter is a specified value θ. Throughout it is assumed that the groups are ordered such that the expected values of Z_t and B_t, and thus θ, are positive under H_1.

For example, suppose that the study design provides power $1 - \beta_D = 0.85$ using a Mantel logrank test at $\alpha_D = 0.05$ (two-sided) to detect a relative hazard of $\phi_D = 0.6$, or equivalently a 40 per cent risk (hazard) reduction, with expected hazard rate of 0.35/year in the control group. From Lachin and Foulkes [12], allowing for 5 per cent losses to follow-up per year, a total $N = 355$ is required for a study with a one year recruitment period and 2.5 year duration total. This design yields a drift parameter of $\theta_D = Z_{0.975} + Z_{0.85} = 1.96 + 1.04 = 3$ (approximately). Under this design one expects $E(D_c) = 85$ events in the control group.

For a futility analysis at $t = 0.5$ ($d_c = 42.5$ or 43), Table I(A) presents the distribution of the interim Z-values ($Z_{0.5}$) assuming the design effect size $\theta = 3$. From this, the distributions of two-sided p-values, B-values ($B_{0.5}$), drift parameter estimates ($\hat{\theta}_{0.5}$) and conditional power values are derived. There is a 50 per cent chance that the nominal two-sided interim p-value > 0.034, and only a 20 per cent chance that it would be > 0.20. Likewise, there is a 50 per cent chance that the drift parameter estimate $\hat{\theta} \geqslant 3.0$. However, if the true risk reduction is only 20 per cent ($\phi = 0.8$), then from Reference [12] the above study plan yields $Z_{1-\beta} = -0.56$ with corresponding power $1 - \beta = 0.29$ and drift parameter $\theta = 1.96 - 0.56 = 1.40$. Table I(B)

Table I. Percentiles of the distribution of the Z-values at $t=0.5$ ($Z_{0.5}$) obtained from (6) assuming $\phi=0.6$ for which the design provides 85 per cent power and $\theta=3$ (A), and assuming $\phi=0.8$ for which $\theta=1.4$ (B).

Percentile	$Z_{0.5}$	p-value	$B_{0.5}$	$\hat{\theta}_{0.5}$	CP_{design}	CP_{trend}	CP_{null}
A. $\phi=0.6$, $\theta=3$							
0.05	0.48	0.63	0.34	0.68	0.430	0.035	0.011
0.2	1.28	0.20	0.91	1.81	0.735	0.416	0.068
0.5	2.12	0.034	1.50	3.00	0.929	0.930	0.26
0.8	2.96	0.0031	2.10	4.19	0.990	0.999	0.58
0.95	3.77	0.0002	2.66	5.33	0.999	1.0	0.84
B. $\phi=0.8$, $\theta=1.40$							
0.05	−0.66	1	−0.46	−0.93	0.095	<0.001	<0.001
0.2	0.15	0.89	0.10	0.21	0.306	0.007	0.004
0.5	0.98	0.33	0.70	1.40	0.631	0.21	0.04
0.8	1.83	0.068	1.29	2.59	0.880	0.81	0.17
0.95	2.63	0.009	1.86	3.72	0.976	0.999	0.44

The corresponding two-sided p-values, B-values ($B_{0.5}$), drift parameter estimates ($\hat{\theta}_{0.5}$) and conditional power values estimated under the original design ($\theta_F=3$), the current trend ($\theta_F=\hat{\theta}_{0.5}$) and under the null hypothesis ($\theta_F=0$) are also presented.

describes the resulting probability distribution of the Z-test value and related quantities at 43 control events ($t=0.5$). Now there is a 50 per cent chance that the interim p-value is >0.33, 20 per cent chance that it is >0.89. There is only a 13 per cent chance that $\hat{\theta}\geqslant 3.0$. Note that t is defined on the basis of the hazard rate and number of events in the control group that is the same for any ϕ.

3. DISTRIBUTION OF CONDITIONAL POWER VALUES

Let θ_I denote the drift parameter for the distribution of the initial observed data that leads to the interim B-value B_t at time t. Then, let θ_F denote the assumed drift parameter for the distribution of the future data. For given θ_F, since the final result at the planned study end is $Z_I=B_1=B_t+U_t$, from the distribution of U_t in Reference [6] conditioned on the observed data B_t, it follows that

$$B_1|(B_t,\theta_F)\sim N(\tilde{\theta},1-t) \tag{7}$$

where

$$\tilde{\theta}=B_t+\theta_F(1-t)=t\hat{\theta}_t+\theta_F(1-t) \tag{8}$$

Then from the standard expression for the power of a Z-test (cf. Reference [13]), it follows that the conditional power is obtained as $CP(t,\theta_F)=\Phi[Z_{CP}(t,\theta_F)]$ where

$$Z_{CP}(t,\theta_F)=\frac{\tilde{\theta}-Z_{1-\alpha_D}}{\sqrt{1-t}}=\frac{B_t+\theta_F(1-t)-Z_{1-\alpha_D}}{\sqrt{1-t}}=\frac{t\hat{\theta}_t+\theta_F(1-t)-Z_{1-\alpha_D}}{\sqrt{1-t}} \tag{9}$$

For an interim analysis at time t, the distribution of B_t is provided by (6) using drift parameter value θ_I. This then provides the unconditional distribution of $Z_{CP}(t, \theta_F)$ for the assumed θ_F:

$$Z_{CP}(t, \theta_F) \sim N\left[\frac{t\theta_I + (1-t)\theta_F - Z_{1-\alpha_D}}{\sqrt{1-t}}, \frac{t}{1-t}\right] \tag{10}$$

that is a function of θ_I as well as θ_F. Herein, computations of conditional power are considered under the original design, or the current trend in the data, or the null hypothesis.

The conditional power under the original design, designated as CP_D, is computed assuming that the future data will be generated as specified in the initial study design where $\theta_F = \theta_D$ is the design-specified drift parameter that provides the desired level of power $1 - \beta_D$ to detect the specified treatment effect ϕ_D, such as $\phi_D = 0.6$ and $\theta_D = 3$ for the above example with $N = 355$. For an analysis at information time t, the conditional power is then derived upon substituting $\theta_F = \theta_D = 3$ into (9), with the resulting distribution given in (10). The distribution of conditional power values could also be determined for any hypothesized value of the true drift parameter θ_I, such as when the true effect size ϕ is smaller than ϕ_D in which case $\theta_I < \theta_D$.

Another approach would be to compute the conditional power under the current trend in the data, designated as CP_T, assuming that the future data will arise from the same distribution that generated the observed data. This computation would employ the interim drift parameter estimate $\theta_F = \hat{\theta}_t$ at time t. In this case the expression for the conditional power in (9) simplifies to

$$Z_{CP}(t, \hat{\theta}_t) = \frac{\hat{\theta} - Z_{1-\alpha_D}}{\sqrt{1-t}} = \frac{B_t/t - Z_{1-\alpha_D}}{\sqrt{1-t}} \tag{11}$$

with unconditional distribution

$$Z_{1-\beta}(t, \hat{\theta}_t) \sim N\left[\frac{\theta_I - Z_{1-\alpha_D}}{\sqrt{1-t}}, \frac{1}{t(1-t)}\right] \tag{12}$$

Alternately, the conditional power under the null hypothesis assuming $\theta_F = 0$, designated as CP_N, is derived from

$$Z_{CP}(t, \theta_F = 0) = \frac{t\hat{\theta}_t - Z_{1-\alpha_D}}{\sqrt{1-t}} = \frac{B_t - Z_{1-\alpha_D}}{\sqrt{1-t}} \tag{13}$$

with unconditional distribution

$$Z_{CP}(t, \theta_F = 0) \sim N\left[\frac{t\theta_I - Z_{1-\alpha_D}}{\sqrt{1-t}}, \frac{t}{1-t}\right] \tag{14}$$

Note that this applies to the distribution of CP_N when the observed data are drawn from a distribution with possibly non-zero drift parameter θ_I and it is assumed that the null hypothesis applies to the future data.

Table I also describes the distribution of the conditional power values CP_D, CP_T, and CP_N for the case where the $\phi = 0.6$ and the design provides $\theta_I = \theta_D = 3$, and that where $\phi = 0.8$ and the design provides $\theta_I = 1.4$. When the true alternative is closer to the null hypothesis, as when $\phi = 0.8$, the conditional power values are lower because the interim Z, B and $\hat{\theta}$ values are likely closer to the null value. Thus, as ϕ approaches the null, the distribution of CP values is shifted towards 0.

4. PROBABILITY OF STOPPING FOR FUTILITY

Many studies include a provision for stopping for futility based on a low conditional power value. Consider the case of a single pre-specified look at time $t=\tau$, with the rule that the study will be stopped for futility if $\mathrm{CP} \leqslant C_L$, with corresponding normal deviate $Z_L = \Phi^{-1}(C_L) = \mathrm{Probit}(C_L)$. The probability of such stopping is

$$P_L = P[\mathrm{CP}(\tau, \theta_\mathrm{F}) \leqslant C_L] = P[Z_{\mathrm{CP}}(\tau, \theta_\mathrm{F}) \leqslant Z_L] = \Phi\left[\frac{Z_L - \mu_z}{\sqrt{\sigma_z}}\right] \tag{15}$$

with respect to the mean μ_z and variance σ_z^2 of the distribution of the CP value at time τ on which basis the study is to be evaluated, such as CP_D, CP_T, or CP_N described above.

Since the value of $Z_{\mathrm{CP}}(\tau)$ is a function of the B-value, the corresponding critical B-value (B_L) when the stopping rule employs CP_D, CP_T, or CP_N is

θ_F	$B_L : Z_\tau = Z_L$
Design (θ_D)	$Z_L\sqrt{1-\tau} - \theta_D(1-\tau) + Z_{1-\alpha_D}$
Trend ($\hat{\theta}_\tau$)	$[Z_L\sqrt{1-\tau} + Z_{1-\alpha_D}]\tau$
Null (0)	$Z_L\sqrt{1-\tau} + Z_{1-\alpha_D}$

(16)

The study is stopped for futility if $B_\tau \leqslant B_L$, or equivalently if $\hat{\theta}_\tau \leqslant \theta_L = B_L/\tau$. It follows that any futility stopping boundary on CP_D, CP_T or CP_N, or some other computation, implies an equivalent boundary on the B-value or interim drift estimate $\hat{\theta}$.

Figure 1 presents the CP values over the range 0.05–0.3 corresponding to B-values for an analysis at $t=0.5$ and 0.8 assuming a drift parameter $\theta=3$. Since $\hat{\theta}_t = B_t/t$ a similar equivalence applies to $\hat{\theta}_t$ as also shown in the figure. The CP function declines sharply over the range of B-values (or $\hat{\theta}$), less so for larger t. Smaller B or $\hat{\theta}$ values are required to yield low CP values under the original design, than under the current trend or the null hypothesis, respectively.

Thus, a boundary for any CP computation implies a boundary for the B or $\hat{\theta}$ values, and the properties of the plan then depend explicitly on the latter. Thus, rather than specifying a futility stopping boundary in terms of a CP computation, it is sufficient to specify that the study would be stopped if $B_\tau \leqslant B_L$, or equivalently if $\hat{\theta}_\tau \leqslant \theta_L$ for an analysis at time τ. For a specified value of $\theta = \theta_\mathrm{I}$ for the observed data, the probability of such stopping (P_L) can then be computed from the distribution of B_t or $\hat{\theta}_t$ in (6) as

$$\begin{aligned} P_L &= P[B_\tau \leqslant B_L \mid \theta_\mathrm{I}] = \Phi\left[\frac{B_L - \tau\theta_\mathrm{I}}{\sqrt{\tau}}\right] \\ &= P[\hat{\theta}_\tau \leqslant \theta_L \mid \theta_\mathrm{I}] = \Phi[(\theta_L - \theta_\mathrm{I})\sqrt{\tau}] \end{aligned} \tag{17}$$

The value computed under H_0: $\theta_\mathrm{I}=0$ is denoted as P_{L0}, under the alternative ($\theta_\mathrm{I} \neq 0$) as P_{L1}.

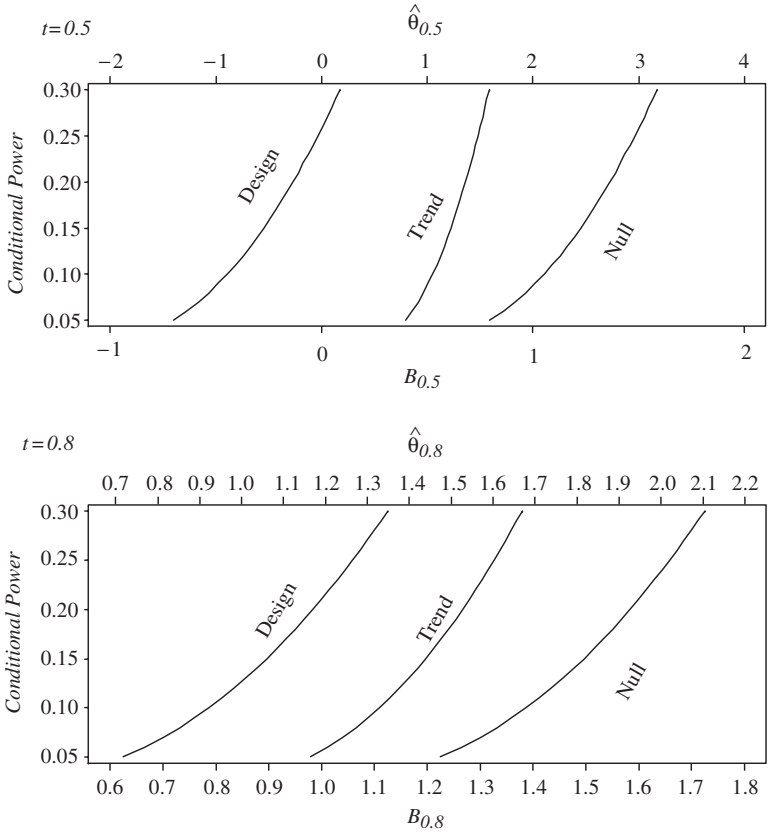

Figure 1. Conditional power values computed under the design ($\theta = 3$), current trend and null hypothesis as a function of the interim B-value B_t and the corresponding drift parameter estimate $\hat{\theta}_t = B_t/t$ at $t = 0.5$ and 0.8.

Figure 2 shows the probability of stopping for futility under the null hypothesis and the design alternative ($\theta = 3$) as a function of the bound on $\hat{\theta}_\tau$ at $\tau = 0.5$ or 0.8 over the range of values θ_L corresponding to $\text{CP}_D = 0.05$ up to $\text{CP}_N = 0.3$. These conditional power levels are achieved over a larger range of values $\hat{\theta}_\tau$ for $\tau = 0.5$ than 0.8. If the true effect size is smaller than that assumed under the design ($\theta < 3$), then the probabilities of stopping are increased.

5. TYPE I AND II ERROR PROBABILITIES

Expressing the futility stopping rule in terms of the interim B-value, or drift estimate $\hat{\theta}$, also facilitates the evaluation of the type I and II error probabilities of the stopping rule. First consider the type II error.

Lan *et al.* [2] present bounds on the type II error probability when a study is terminated due to low conditional power CP_D computed under the original design. If the interim results

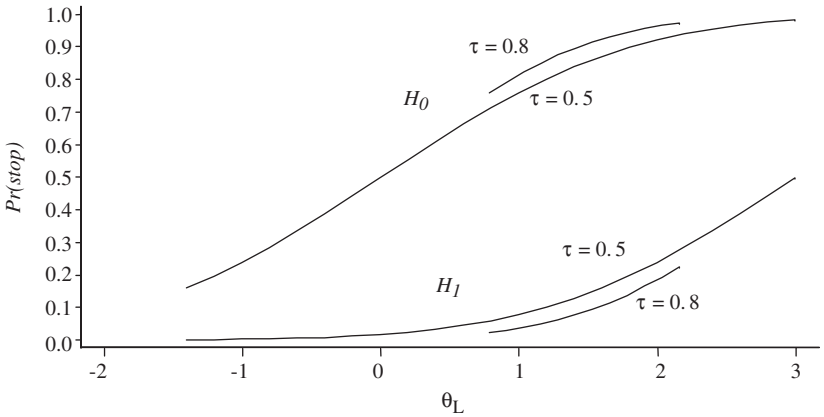

Figure 2. Probabilities of stopping for futility at $\tau = 0.5$ or 0.8 over a range of stopping boundary values of $\theta_{(\tau)L}$ under the null hypothesis $H_0: \theta = 0$ and under the alternative $H_1: \theta = 3$.

are monitored continuously with the plan to terminate the trial under a pre-defined stopping rule $\mathrm{CP}_{Dt} \leqslant \gamma$ at any time t, then the type II error is bounded by $\beta \leqslant \beta_D/(1-\gamma)$.

However, the exact type II error probability can be computed directly for a futility stopping rule at a pre-specified time τ. In this case, if the study is stopped for futility, then under H_1 a type II error is committed. Thus, the total type II error probability is partitioned as $\beta = \beta_1 + \beta_2$ where $\beta_1 = P_{L1}$ is the probability of stopping for futility (and of a type II error) at the pre-specified interim analysis, and β_2 is the probability of continuation and non-significance at the final analysis. The latter probability is

$$\begin{aligned}\beta_2 &= P(B_1 < Z_{1-\alpha_D} \cap B_\tau > B_L) \\ &= P[\{U_\tau < Z_{1-\alpha_D} - B_\tau \mid \theta_{\mathrm{F}}\} \cap \{B_\tau > B_L \mid \theta_{\mathrm{I}}\}] \\ &= \int_{b=B_L}^{\infty} f(b;\theta_{\mathrm{I}}) \int_{u=-\infty}^{Z_{1-\alpha_D}-b} g(u;\theta_{\mathrm{F}})\,\mathrm{d}u\,\mathrm{d}b \end{aligned} \tag{18}$$

where $f(b;\theta_{\mathrm{I}})$ is the distribution of B_τ and $g(u;\theta_{\mathrm{F}})$ that for U_τ as in (6) with respect to the drift parameters assumed for the initial and future data. The value B_L is the lower limit on the value of B_τ for which the study would be continued. For a two-sided test, an exact computation would employ integration of u over $(Z_{\alpha_D/2} - b,\ Z_{1-\alpha_D/2} - b)$. However, it is usually sufficient to simply substitute $Z_{1-\alpha_D/2}$ for $Z_{1-\alpha_D}$ in (18) since the mass over $(-\infty,\ Z_{\alpha_D/2} - b)$ is virtually zero for θ positive away from zero.

Note that

$$\begin{aligned}\beta = \beta_1 + \beta_2 &= P(B_\tau \leqslant B_L) + P(B_1 < Z_{1-\alpha_D} \cap B_\tau > B_L) \\ &< P(B_\tau \leqslant B_L) + P(B_1 < Z_{1-\alpha_D}) = \beta_1 + \beta_D \end{aligned} \tag{19}$$

and $0 \leqslant \beta - \beta_D < \beta_1 = P_{L1}$. Thus, the inflation in the total type II error probability is bounded by the probability of stopping for futility.

For example, the study design employed above provides unconditional power = 0.85 or $\beta_D = 0.15$ when $\phi = 0.6$. Assume that it was pre-specified that the study would be stopped for futility at $\tau = 0.5$ if $\mathrm{CP}_{D(0.5)} \leqslant 0.3$, or equivalently if $B_{0.5} \leqslant B_L = 0.08916$ or $\hat{\theta}_{0.5} \leqslant \theta_L = 0.17831$. If the data were monitored continuously, rather than at a single pre-specified point in time, then the Lan, Simon Halperin bound specifies that the type II error probability is bounded by $0.15/0.7 = 0.214$. However, for a single assessment at $\tau = 0.5$, $\Pr(\mathrm{CP}_{D(0.5)} \leqslant 0.3) = 0.023$. Thus, the total type II error probability is

$$\begin{aligned}\beta &= \Pr(\mathrm{CP}_{D(0.5)} \leqslant 0.3) + \Pr[(|\, Z_1 \,| < 1.96 \,|\, \theta_D) \cap (\mathrm{CP}_{D(0.5)} > 0.3)] \\ &= 0.023 + 0.130 = 0.153. \end{aligned} \tag{20}$$

In this expression, note that $\Pr[|\, Z_{1D} \,| < 1.96] \cong \Pr[Z_{1D} < 1.96]$ since the expectation under the alternative is assumed to be positive by construction.

Figure 3 shows the type II error probabilities associated with a stopping boundary $\hat{\theta}_\tau \leqslant \theta_L$ under the assumption that the specified design effect applies, or $\theta_\mathrm{I} = \theta_\mathrm{F} = \theta_D = 3$. As the boundary increases, the probability of stopping increases and thus the probability of a type II error increases. Since the boundary value for $\hat{\theta}_\tau$ using conditional power under the original design (CP_D) is less than that under the current trend (CP_T), which is less than that under the null (CP_N) (see Figure 1), the inflation in β is greater using CP_T and even greater using CP_N. In some cases this inflation is substantial. For example, $\beta = 0.1505$ when stopping for futility using $\mathrm{CP}_{D(0.5)} \leqslant 0.2$ or $\hat{\theta}_{0.5} \leqslant -0.270$. However, stopping when $\mathrm{CP}_{T(0.5)} \leqslant 0.2$ or $\hat{\theta}_{0.5} \leqslant 1.36$ results in $\beta = 0.203$, and stopping when $\mathrm{CP}_{N(0.5)} \leqslant 0.2$ or $\hat{\theta}_{0.5} \leqslant 2.73$ results in $\beta = 0.442$.

In a similar manner the probability of a type I error can be computed when the study plan allows termination for futility. In this case, the type I error probability will be reduced since there is a chance that the study will be stopped for futility. For any conditional power computation (e.g. CP_D or CP_T or CP_N), the probability of stopping for futility is computed under the null hypothesis H_0: $\theta = 0$; i.e. using $\theta_\mathrm{I} = 0$, with boundary Z_L in (15) or B_L (16).

Again, the total type I error probability is partitioned as the sum of the errors when the study is stopped for futility or is continued, $\alpha = \alpha_1 + \alpha_2$, where $\alpha_1 = 0$. Thus, $\alpha = \alpha_2$ is the probability of continuation and significance at the final analysis under H_0: $\theta = 0$.

For a one-sided test at level α_D specified in the study design, the resulting α_2 is

$$\begin{aligned}\alpha_2 &= P(B_1 \geqslant Z_{1-\alpha_D} \cap B_t > B_L \,|\, H_0) \\ &= P[\{U_t \geqslant Z_{1-\alpha_D} - B_t\} \cap \{B_t > B_L\} \,|\, H_0] \\ &= \int_{b=B_L}^{\infty} f_0(b) \int_{u=Z_{1-\alpha_D}-b}^{\infty} g_0(u)\, \mathrm{d}u\, \mathrm{d}b \end{aligned} \tag{21}$$

where $f_0(b)$ is the distribution of B_t and $g_0(u)$ that for U_t as in (6) each evaluated under the null hypothesis with $\theta = 0$. The probability for a two-sided test is twice that above using $Z_{1-\alpha_D/2}$.

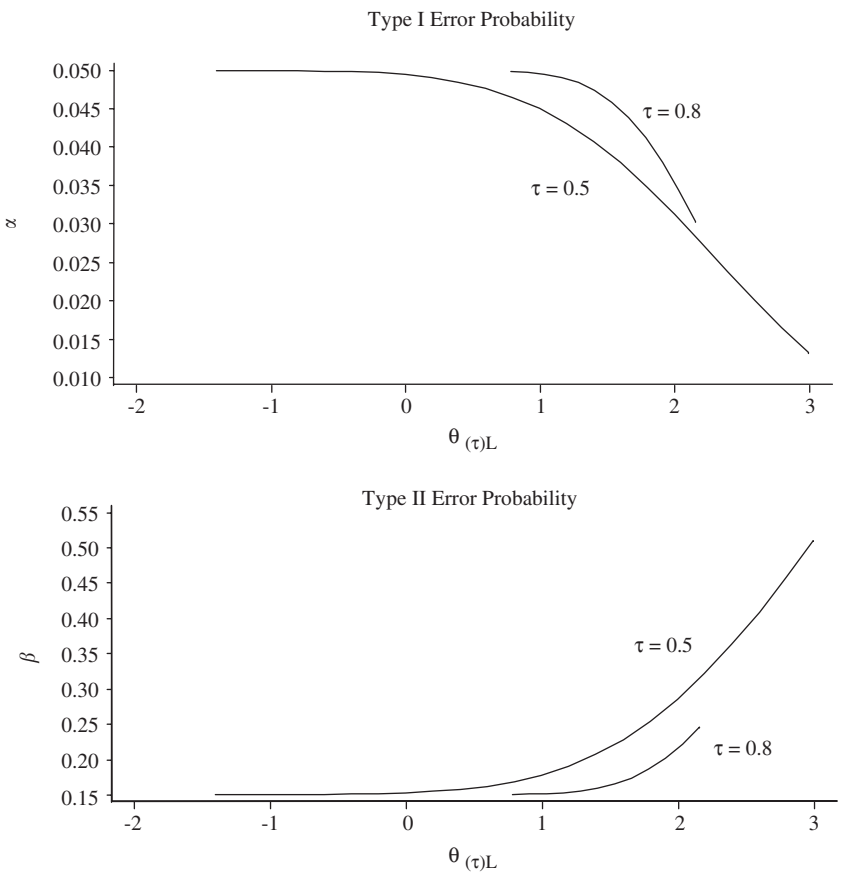

Figure 3. Type I error probability (α) under the null hypothesis $H_0: \theta = 0$, and Type II error probability (β) under the alternative hypothesis $H_1: \theta = 3$ when stopping for futility at $\tau = 0.5$ or 0.8 over a range of stopping boundary values of $\theta_{(\tau)L}$.

Figure 3 also shows the type I error probabilities associated with a stopping boundary defined in terms of θ_L. As the boundary and the probability of stopping decrease, the probability of a type I error approaches the non-sequential level α.

6. FIXING THE ERROR PROBABILITIES

Since the probabilities of type I and II errors are smooth functions of the critical value for the final test and the futility stopping boundary B_L or θ_L, it is usually possible to determine precisely the values for which desired error probabilities are obtained. For example, given a stopping boundary θ_L, the critical value, say Z_F, can be determined such that a test at level α will provide desired type I error probability α_D for the trial. This is obtained by iterative solution of (21) such that $\alpha_2 = \alpha_D$ upon substituting Z_F for $Z_{1-\alpha_D}$. For example, for a stopping

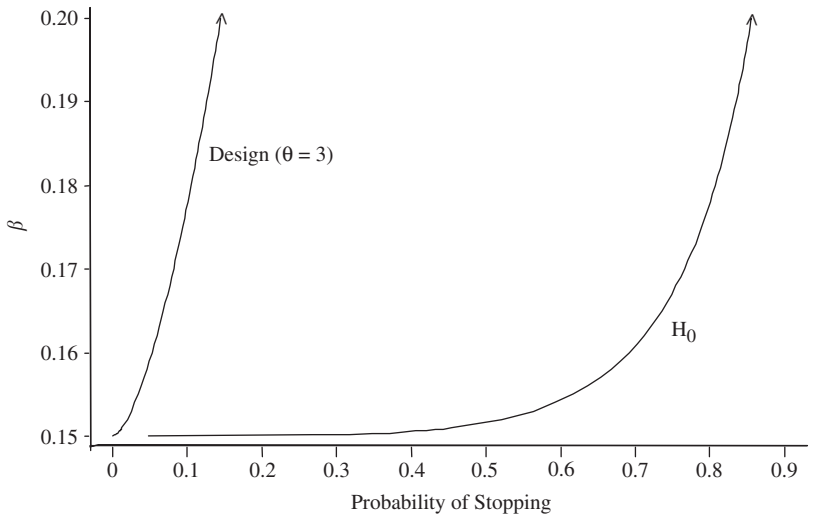

Figure 4. Total type II error probability β as a function of the probability of stopping under H_0 and under $H_1: \theta = 3$ where the type I error probability is controlled at $\alpha = 0.05$ (two-sided).

rule with $\theta_L = 1.8$ at $\tau = 0.5$, $Z_F = 1.7535$ provides $\alpha = \alpha_D = 0.05$ two-sided. The greater the probability of stopping, the smaller the value Z_F required.

The use of a smaller critical value $Z_F < Z_{1-\alpha_D}$ in the final analysis also provides the opportunity to regain power and reduce or eliminate the inflation in the type II error probability. For example, consider the case where the non-sequential study design provides error probabilities α_D and β_D with drift parameter θ_D. Then a futility stopping boundary θ_L and test critical value Z_F are desired such that $\alpha = \alpha_D$ and the type II error probability equals a specified value $\beta \geqslant \beta_D$ for an interim analysis at time τ. The value Z_F is obtained by solution of (21) for a given value of $B_L = \theta_L * \tau$, and the value of B_L by solution of (15) and (18) such that $\beta = \beta_1 + \beta_2$ for a fixed value of Z_F, where Z_F is substituted for $Z_{1-\alpha_D}$. The joint iterative solution proceeds until the values (Z_F, B_L) converge to constants. In such computations, the critical value Z_F is a near-linear function of the realized β, approaching the design $Z_{1-\alpha_D}$ as the specified $\beta \downarrow \beta_D$.

For example, for a futility analysis at $\tau = 0.5$ with $\theta_D = 3$, $\alpha_D = 0.05$ (two-sided), $\beta_D = 0.15$ and $\beta = 0.20$, a 1/3rd increase, joint iterative solution yields $Z_F = 1.8356$ and $B_L = 0.7505$ for which $\theta_L = 1.501$. The probability of stopping for futility under the alternative hypothesis $(H_1: \theta = 3)$ is $P_{L1} = 0.145$ whereas that under the null hypothesis is $P_{L0} = 0.856$. For a smaller 1/6th increase in $\beta = 0.175$, the critical values are $B_L = 0.5673$ or $\theta_L = 1.135$ and $Z_F = 1.8954$ with stopping probabilities of 0.789 and 0.094 under H_0 and H_1, respectively.

Figure 4 shows the relationship between the Type II error probability and the probability of stopping per se for a stopping rule at $\tau = 0.5$ where the value Z_F is simultaneously determined so as to maintain $\alpha = 0.05$, two-sided. As the probability of stopping under H_0 decreases, the inflation in β decreases to trivial values. For example, the boundary $B_L = 0$ allows $P_{L0} = 0.50$ by definition. To maintain $\alpha = 0.05$, two-sided, $Z_F = 1.5451$ for which $\beta = 0.15162$, only trivially inflated. Thus, with this approach it is possible to completely eliminate the inflation in

type II error probability by choice of the critical value Z_F provided that a low probability of stopping for futility is acceptable. For an analysis at $\tau = 0.5$, the boundary $B_L = -1.1794$, or equivalently $\theta_L = -2.359$, when used with $Z_F = 1.95996$ yields $\alpha = 0.05$ (two-sided) and $\beta = 0.15$, with probabilities of stopping of 0.000076 under the design assumptions and only 0.047 under the null.

The fundamental issue is that there is a direct relationship between the probability of stopping under the null hypothesis and the inflation in the type II error probability. Thus, an alternate approach is to first determine the boundary B_L from Equation (17) that provides a specified probability of stopping P_{L0} under the null, and then determine the critical value Z_F and the resulting type II error probability β. For example, at $\tau = 0.5$, $B_L = 0.47692$ provides stopping probability $P_{L0} = 0.75$ under the null hypothesis. With this boundary, iterative solution yields $Z_F = 1.91413$ for type I error probability $\alpha = 0.05$ (two-sided). If the design provides $\theta_D = 3$ or $\beta_D = 0.15$, then these values of B_L and Z_F yield $\beta = 0.16719$ and stopping probability $P_{L1} = 0.074$.

It should be noted, however, that the inflation in type II error probability is markedly reduced when the futility stopping rule applies later in the study, such as at $\tau = 0.8$. In this case, even for a probability of stopping under H_0 as high as 0.85, the inflation would be to $\beta = 0.1513$ when $\beta_D = 0.15$.

7. FUTILITY STOPPING AND ALPHA SPENDING

The expressions herein are based on a single interim analysis at a pre-specified time with no other interim analysis plans. In many instances, however, a conditional power assessment for futility is employed in addition to a group sequential plan to monitor effectiveness, such as using the Lan–DeMets [5] α-spending function. The Reboussin *et al.* [14] computer program, among others, will compute group sequential boundaries and will also compute boundary crossing probabilities for a given value of the drift parameter θ. This allows the implementation of group sequential monitoring for effectiveness in conjunction with a provision for futility stopping.

If the futility monitoring plan specifies that the futility assessment be conducted at $\tau = 0.5$, then in many cases it would be acceptable to forego group sequential analysis until that time since it is highly unlikely that the group sequential boundary would be crossed prior to that time. In this case, the alpha spending could begin at that or the next analysis. Rather than spending the original design α_D over the remainder of the trial, one need only spend α_F, or the probability corresponding to the critical value Z_F. In this case the above calculations would still apply approximately when using an O'Brien–Fleming-type spending function for which the realized group sequential power remains very close to the non-sequential power. With a Pococok-type spending function for which power is reduced, a direct calculation is required as described by Strömberg *et al.* [15] in order to maintain the desired level of power.

For example, for a single interim analysis at $\tau = 0.5$, it is shown above that futility stopping with $B_L = 0.7505$ (or $Z_L = 1.0614$) and a final $Z_F = 1.8356$, provides $\alpha = 0.05$ (two-sided) and $\beta = 0.20$ when $\beta_D = 0.15$. The corresponding one-sided type 1 error probability at the final look, conditional on continuation, is $\alpha_F = 1 - \Phi(Z_F) = 0.03321$. The upper one-sided O'Brien–Fleming-type bounds at $t = 0.5$ and 1.0 computed by the Reboussin *et al.* [14] program with type I error probability α_F are 2.7946 and 1.8470, respectively. The program can then be used

to compute the 'exit' probabilities with these upper bounds and the lower bounds 1.0614 and -10.0, the latter to specify that there is no futility bound in the final analysis. Under H_0, the cumulative exit probability is 0.88105 that includes the probability of stopping for futility $P_{L0} = 0.85574$ so that the total $\alpha = 0.88105 - 0.85574 = 0.02531$. Another calculation using these bounds with drift parameter $\theta = 3$ yields a cumulative exit probability 0.94481. Noting that $P_{L1} = 0.14459$ the total power is $1 - \beta = 0.94481 - 0.14459 = 0.80022$. Thus, this simple approach provides both an upper bound for effectiveness and a single lower bound for futility with nearly the desired error probabilities. The lower the probability of stopping for futility, the more exact the error probability computations.

Such calculations with an O'Brien–Fleming-type α-spending function could be used for any number of analyses beyond the futility stopping time because the power is largely unaffected by the number and timing of the analyses. The lower boundary would be -10 for all but the first analysis.

When group sequential monitoring is also employed prior to the specified time of the futility analysis, the error probabilities of the futility monitoring plan will be affected, and likewise the futility bound will affect the properties of the group sequential plan for subsequent analyses. In general, if K group sequential analyses are conducted prior to the futility analysis, then the distribution of the interim Z or B-values at the $(K+1)$th analysis at time τ is obtained from the K-fold convolution of the distributions of the independent increments in the test statistic, integrated with respect to the group-sequential continuation region, as described by Armitage, McPherson and Rowe [16], among others. That distribution, obtained by numerical integration, would then be substituted for the simple normal distributions in (6) and employed in the integrals used to compute the probability of stopping and the error probabilities α and β.

However, the effects are trivial with an O'Brien–Fleming-type boundary and the iterative procedure herein can be used along with the Reboussin *et al.* [14] program to determine both a futility stopping boundary and an α-spending boundary with approximately the desired levels α and β. The computations are best described by example. For group sequential analyses for effectiveness starting at $t = 0.25$ with increments of 0.125, in conjunction with a provision for stopping for futility at $\tau = 0.5$, the following upper and lower (futility) bounds

t	0.25	0.325	0.5	0.625	0.75	0.8785	1
Z_U	4.3326	3.4814	2.8006	2.5013	2.2725	2.0963	1.9554
Z_L	-	-	1.06427	-	-	-	-

yield total $\alpha = 0.0266$ one sided (0.0532 two-sided) and $\beta = 0.2043$. These boundaries were computed as follows.

The upper α-spending boundaries were computed at the first two analyses (prior to the futility assessment) for a plan with total $\alpha = 0.025$ (one-sided). The cumulative α spent after the second look is 0.00025 and the remaining α to be spent is 0.02475. Then, for an interim futility analysis at $\tau = 0.5$ with $\alpha_D = 0.02475$, $\beta_D = 0.15$, and total $\beta = 0.20$ for $\theta = 3$, the above iterative procedure yields $B_L = 0.75255$ or $Z_L = 1.06427$ and a final test critical value $Z_F = 1.84009$ with nominal $\alpha_F = 0.032877$. The remaining α-spending upper boundary values are then computed with one-sided total $\alpha = \alpha_F$ for analyses at $t = 0.5$, 0.625, 0.75, 0.875 and 1.0.

To compute the total error probabilities the Reboussin *et al.* program [14] is used with the above upper and lower boundaries, using -10 in place of '-', to compute cumulative exit probabilities of 0.88301 under H_0 and 0.94040 under $H_1: \theta = 3$. These include the probabilities of stopping for futility. The program is then used with the boundaries

t	0.25	0.325	0.5
Z_U	4.3326	3.4814	10
Z_L	-10	-10	1.06427

to compute the exit probabilities of stopping for futility at $t = 0.5$ of 0.85640 under H_0 and 0.14524 under H_1. Then the total $\alpha = 0.88301 - 0.85640 = 0.02661$ and power $1 - \beta = 0.94040 - 0.14524 = 0.79516$. The type I error probability is slightly inflated whereas the power is unaffected.

In many cases, this type of computation will satisfactorily control both type I and II error probabilities. If desired, the bounds beyond the futility analysis could be obtained using a slightly smaller value $\alpha < \alpha_F$ so that the type I error probability would be no greater than the desired 0.025 with a negligible loss in power.

If the futility rule is applied later in the study, this approach may satisfactorily preserve the desired type I error probability. For example, if bounds are computed with a futility assessment at $\tau = 0.75$ rather than at 0.5, then the realized $\alpha = 0.02461$ and power $1 - \beta = 0.7817$.

The above approach does not generalize to more than two looks, although in principle upper and lower boundary values could be determined with a provision for futility stopping only at specific looks, such as at $t = (0.75, 0.875)$ above, with specific error probabilities α and β. Other approaches that meld group sequential monitoring for effectiveness with sequential monitoring for futility are described below.

8. DISCUSSION

Conditional power is a useful concept for the consideration of early termination for futility and is a concept that is readily understood by clinical investigators. However, there is no unique CP value because there is an infinite range of possible assumptions or values θ_F regarding the distribution of the future data. Thus, the CP computed under the initial design differs from that under the current trend or the null hypothesis or any other value θ_F assumed for the future data. Usually a futility stopping rule consists of a specified lower boundary for a specific computation, such as the CP under the design. Such a specification also provides specific boundaries in terms of the observed interim data represented by the current Z-value Z_t, or the corresponding B-value B_t, or the equivalent drift parameter estimate $\hat{\theta}_t$. Thus, a boundary for a particular CP value is equivalent to a boundary on B_t and *vice versa*. Further, the properties of the stopping plan including the probabilities of stopping under the null and alternative hypotheses and the actual type I and II error probabilities are explicit functions of Z_t, B_t, or $\hat{\theta}_t$, regardless of the specific CP computation chosen. Thus, it is sufficient to describe a futility stopping rule and its properties in terms of Z_t, B_t or $\hat{\theta}_t$.

Early termination based on conditional power computations, or equivalently the values of B_t or $\hat{\theta}_t$, may substantially inflate the type II error probability β when there is a high probability of early termination. The probability β increases as the probability of stopping increases. This

inflation can be mitigated by iterative solution of a smaller critical value for the final test Z_F that provides type I error probability exactly equal to the desired level α_D. However, with high probability of stopping, this inflation in β cannot be eliminated and some modest inflation must be allowed. As the probability of stopping decreases towards zero, the increases in β become trivial. For a probability of 0.5 of stopping under H_0 at $\tau = 0.5$, the type II error probability is inflated from $\beta = 0.15$ to only 0.1516.

The procedures herein allow one to determine the futility stopping boundary B_L and the critical value Z_F that provide specified error probabilities $\alpha = \alpha_D$ and $\beta \geqslant \beta_D$, from which the associated probability of stopping can be computed under the null hypothesis or some other sub-optimal effect $0 \leqslant \phi < \phi_D$. Alternately, the value B_L that provides a specified probability of stopping under H_0 or a specified effect size can be determined. Then the critical value Z_F that provides the desired $\alpha = \alpha_D$ and the type II error probability computed can be determined.

Snappin [17] presents expressions for the probabilities of acceptance or rejection when the stopping rule specifies both a criterion for rejection of H_0 based on the current trend in the data and acceptance of H_0 based on the original design. By simulation he also shows that the type I error probability is not inflated whereas the type II error probability can be inflated when using a rule to stop for rejection of H_0. Pepe and Anderson [18] present expressions for the type I and II error probabilities similar to those herein for use with time to event data when the conditional power is computed under a conservative estimate of the effect size ϕ. Like Ware *et al.* [19] they show that there is minimal inflation in the type II error probability when stopping for futility based on CP_D or a more conservative projection. They also describe an iterative procedure to determine the critical Z-value and total N such that the desired error probability levels were obtained for their fixed stopping rule. However, these papers do not consider the effect on the error probabilities using conditional power values evaluated under other less conservative projections for the future data such as the null hypothesis or the current trend in the data. Such procedures provide a higher probability of stopping and thus a higher probability of a type II error compared to conditional power calculations under the original, or a reasonable, design.

This inflation in type II error probabilities also applies to some procedures for sample size re-estimation during a trial that are based on conditional power computed under the current trend, such as the procedure of Lan and Trost [20]. Such procedures guarantee the conditional power of the study, conditioned on not stopping for futility. However, the *unconditional* (overall) type II error probability then consists of the probability of stopping for futility plus the probability of continuing the study (possibly after resizing) and then reaching a negative result. Even though the conditional power may be guaranteed (say at 0.85), the total type II error probability will be inflated (beyond 0.15 in this case) due to the additional probability of stopping for futility even though the alternative hypothesis is true.

For a survival analysis, Ellenberg and Eisenberger [21] and Wieand *et al.* [22] suggest that termination for futility be considered when the relative risk estimate from a proportional hazards model is less than 1, which is equivalent to a rule for termination when $B_\tau \leqslant 0$. As stated above, for an analysis at $\tau = 0.5$, $\alpha = 0.05$ is provided by $Z_F = 1.95451$ rather than the nominal 1.96. Using this critical value for a study with $\theta_D = 3$ and $\beta_D = 0.15$, the type II error probability is $\beta = 0.15162$, only slightly inflated. This also provides a probability of stopping of 0.50 under H_0 *versus* 0.017 under H_1. This approach is advantageous in terms of the impact on the error probabilities provided that the 50 per cent chance of stopping under H_0 is acceptable.

While the methods herein have principally focused on the logrank test for survival data under the proportional hazards model, Lin *et al*. [23] describe a general theory for the assessment of conditional power for the family of weighted logrank tests, that includes among others, the variations of the Wilcoxon test, possibly with covariance adjustment.

The methods herein provide for only a single futility analysis, possibly in conjunction with a group sequential analysis for effectiveness. Whitehead and Matsushita [24] describe a futility design based on the theory of sequential score statistics that allows continuous or discrete time monitoring for futility over the life of the trial but in a setting with no sequential monitoring for effectiveness. They also describe a comparable procedure based on stochastic curtailing. Friedlin and Korn [25] evaluate the properties of aggressive methods for stopping for futility for event time data, and show that a sequential futility stopping rule is equivalent to a specific lower boundary in the Pampallona and Tsiatis [26] group sequential design with an O'Brien–Fleming like upper bound. Betensky [27] describes a group sequential procedure with an upper boundary for effectiveness, and a lower boundary for futility based on a generalization of the Pepe–Anderson conditional power computation. She also describes an upper bound on the power of the test [28].

As an alternative to using conditional power to monitor futility, numerous authors [26, 29–34] present group sequential procedures that provide bounds for 'acceptance' or rejection of H_0 while preserving the desired type I error probability, and possibly the type II error probability as well. Since these bounds apply to the sequential Z or B-values, the lower bound for such designs can also be described in terms of a bound on the corresponding conditional power. For example, consider a group sequential design with O'Brien–Fleming upper and lower bounds for effectiveness and futility, respectively, that provides power of 0.80 for $\alpha = 0.05$ (two-sided). With equally spaced interim analyses, the lower boundary on the interim Z-value corresponds to a futility stopping boundary for a conditional power $\mathrm{CP}_D < 0.5$ computed under the initial design [33].

All calculations herein were conducted using programs written in SAS that employ numerical integration with a grid width of 0.0001. Programs are available from the author.

ACKNOWLEDGEMENTS

This work was partially supported by a Cooperative Agreement from the National Institute of Diabetes, Digestive and Kidney Diseases for the type I Diabetes TrialNet Co-ordinating Center. The author is grateful for his many discussions with K.K. Gordon Lan on the theory of sequential analysis on which this paper is based. The author also acknowledges helpful discussions with Jim Rochon, Michael Proschan and Naji Younes, and thanks Ms Paula Friedenberg for programming assistance.

REFERENCES

1. Halperin M, Lan KKG, Ware JH, Johnson NJ, DeMets DL. An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials* 1982; **3**:311–323.
2. Lan KKG, Simon R, Halperin M. Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics C* 1982; **1**:207–219.
3. Davis BR, Hardy RJ. Upper bounds for type I and type II error rates in conditional power computations. *Communications in Statistics—Theory and Methods* 1990; **19**:3571–3584.
4. Davis BR, Hardy RJ. Repeated confidence intervals and prediction intervals using stochastic curtailing. *Communications in Statistics—Theory and Methods* 1992; **21**:351–368.
5. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
6. Lan KKG, Wittes J. The *B*-value: a tool for data monitoring. *Biometrics* 1988; **44**:579–585.

7. Lan KKG, Zucker DM. Sequential monitoring of clinical trials: the role of information and Brownian motion. *Statistics in Medicine* 1993; **12**:753–765.
8. Lan KKG, Reboussin DM, DeMets DL. Information and information fractions for design and sequential monitoring of clinical trials. *Communications in Statistics—Theory and Methods* 1994; **23**:403–420.
9. Wu MC, Lan KKG. Sequential monitoring for comparison of changes in a response variable in clinical trials. *Biometrics* 1992; **48**:765–779.
10. Tsiatis AA. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* 1982; **77**:855–861.
11. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* 1981; **2**:93–113.
12. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, non-compliance and stratification. *Biometrics* 1986; **42**:507–519.
13. Lachin JM. *Biostatistical Methods*: *The Assessment of Relative Risks*. Wiley: New York, 2000.
14. Reboussin DM, DeMets DL, Kim K, Lan KKG. Computations for group sequential boundaries using the Lan–DeMets spending function method. *Controlled Clinical Trials* 2000; **21**:190–207.
15. Strömberg U, Losic N, Lanke J. Incorporation of a stopping criterion for futility into the design of a clinical trial with one interim analysis. *Communications in Statistics—Theory and Methods* 1997; **26**:1011–1030.
16. Armitage P, McPherson CK, Rowe BC. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society*, *Series A* 1969; **132**:235–244.
17. Snapinn SM. Monitoring clinical trials with a conditional probability stopping rule. *Statistics in Medicine* 1992; **11**:659–672.
18. Pepe MS, Anderson GL. Two-stage experimental designs: early stopping with a negative result. *Applied Statistics* 1992; **41**:181–190.
19. Ware JH, Muller JE, Braunwald E. The futility index. *American Journal of Medicine* 1985; **78**:635–643.
20. Lan KKG, Trost DC. Estimation of parameters and sample size re-estimation. *Proceedings of the American Statistical Association Biopharmaceutical Section* 1997; 48–51.
21. Ellenberg SS, Eisenberger MA. An efficient design for Phase III studies of combination chemotherapies. *Cancer Treatment Reports* 1985; **69**:1147–1152.
22. Wieand S, Schroeder G, O'Fallon JR. Stopping when the experimental regimen does not appear to help. *Statistics in Medicine* 1994; **13**:1453–1458.
23. Lin DY, Yao Q, Ying Z. A general theory on stochastic curtailment for censored survival data. *Journal of the American Statistical Association* 1999; **94**:510–521.
24. Whitehead J, Matsushita T. Stopping clinical trials because of treatment ineffectiveness: a comparison of a *futility design* with a method of stochastic curtailment. *Statistics in Medicine* 2003; **22**:677–687.
25. Friedlin B, Korn EL. A comment on futility monitoring. *Controlled Clinical Trials* 2002; **23**:355–366.
26. Pampallona S, Tsiatis AA. 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* 1994; **42**:19–35.
27. Betensky RA. Conditional power computations for early acceptance of H_0 embedded in sequential tests. *Statistics in Medicine* 1997; **16**:465–477.
28. Betensky RA. Early stopping to accept H_0 based on conditional power: approximations and comparisons. *Biometrics* 1997; **53**:794–806.
29. DeMets DL, Ware JH. Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* 1982; **69**:661–663.
30. Wang SK, Tsiatis AA. Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 1987; **43**:193–199.
31. Emerson SS, Fleming TR. Symmetric group sequential test designs. *Biometrics* 1989; **45**:905–923.
32. Whitehead J. *The Design and Analysis of Sequential Clinical Trials* (2nd edn). Wiley: Chichester, 1997.
33. Kittleson JM, Emerson SS. A unifying family of group sequential test designs. *Biometrics* 1999; **55**:874–882.
34. Pampallona S, Tsiatis AA, Kim KM. Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Information Journal* 2001; **35**:1113–1121.

