CLINICAL TRIALS ARTICLE Clinical Trials 2010; 0: 1–12

A general inefficacy interim monitoring rule for randomized clinical trials

Boris Freidlin[a], *Edward L Korn*[a] *and Robert Gray*[b]

Background The ultimate goal of a phase III randomized clinical trial designed to demonstrate superiority of a new versus standard therapy is to provide sufficiently compelling evidence to affect clinical practice. To balance patient interests against the need for acquiring evidence it is desirable to stop a study for inefficacy as soon as convincing evidence that the new therapy is not beneficial becomes available.
Purpose To discuss potential deficiencies in some commonly used inefficacy monitoring rules and to propose a comprehensive inefficacy monitoring procedure.
Methods The proposed approach is developed using clinical, logistical, and statistical considerations. The new approach is compared to the commonly used inefficacy rules in a simulation study.
Results Some of the commonly used inefficacy rules are suboptimal with respect to the strength of evidence required for stopping throughout the trial: too conservative in the middle and/or too aggressive at the end. Our approach allows timely stopping (a) if the new therapy is harmful, and (b) if the interim data provides convincing evidence that the new therapy has no tangible benefit. Relative to common inefficacy rules, our procedure is shown to result in potentially fewer treated patients and shorter study duration under the null hypothesis with only a minor loss of power under the alternative hypothesis.
Limitations The proposed procedure is applicable to superiority designs with well-defined clinical objectives.
Conclusions The proposed inefficacy approach is attractive from statistical, clinical, and logistical standpoints. By decreasing average stopping times relative to the commonly used boundaries, our rule lessens patient exposure to inactive treatments, improves resource utilization, and accelerates dissemination of important clinical information. At the same time, the proposed rule provides a clear benchmark for providing compelling evidence that the new therapy is not beneficial. *Clinical Trials* 2010; **0**: 1–12. http://ctj.sagepub.com

Introduction

Randomized clinical trials (RCTs) designed to demonstrate that a new therapy is superior to the current standard need to provide sufficiently compelling evidence to convince the medical community either that the new therapy is beneficial or that the benefit-to-risk ratio of the new therapy does not support its use. To balance the need for sufficient evidence and the need to provide the best available treatment for current and future patients, RCT designs should incorporate interim monitoring of outcome data for early evidence of efficacy or inefficacy [1]. The efficacy interim monitoring allows the study to stop early if sufficiently convincing evidence of the new therapy's superiority is obtained. The inefficacy monitoring covers two contingencies. First, one would need to be able to

[a]Biometric Research Branch, National Cancer Institute, Bethesda, MD, USA, [b]Eastern Cooperative Oncology Group, Harvard School of Public Health, Boston, MA, USA
Author for correspondence: Boris Freidlin, National Cancer Institute, Bethesda, MD, USA.
E-mail: freidlinb@ctep.nci.nih.gov

10.1177/1740774510369019

stop as soon as possible if the new therapy is harmful or is inferior to the control with respect to a relevant clinical outcome (the ethical motivation for this contingency is self-evident). Second, it is desirable for the monitoring to be able to stop the study early once sufficiently convincing evidence that the new therapy does not offer therapeutically tangible benefit is obtained. Motivations for this aspect of inefficacy include minimizing patient exposure to an ineffective and potentially toxic therapy, conserving recourses, informing the clinical community of the results as soon as possible, and making it easier for other therapies to be tested in the same patient population.

Monitoring for inefficacy is frequently called futility monitoring, but we prefer to restrict the use of the word 'futility' to its lay language and original statistical usage [2]: it is futile to continue a trial when, based on the outcome data observed to date, there is a high probability that the trial would not reject the null hypothesis if it continued to its end. If the interim data are convincing that the new therapy does not offer benefits over the standard therapy, then it would be futile to continue the trial. However, unless considered from a purely regulatory perspective, the converse may not be true: the interim data may suggest that it is futile to continue but yet not be convincing that the new therapy is ineffective. This is a result of the asymmetry in the way clinical trials are designed and interpreted. Phase III clinical trial designs contain an explicit definition for the minimal level of evidence required to show that the new therapy is beneficial. At the same time, very few trials are explicitly designed to ensure that failing to reach that minimal level of evidence at the final analysis is sufficiently compelling to prove that the new therapy has no tangible benefit. We argue that an inefficacy boundary should be based on providing sufficiently compelling evidence of lack of benefit and not on the likelihood of rejecting the null hypothesis if the study is continued to its end. We return to this point in the Discussion.

The present work is motivated by our observation that many RCTs reported in the literature employ inefficacy monitoring rules that we consider suboptimal: (a) Some popular rules become too aggressive in the second half of the study and recommend stopping the study for inefficacy at a time when the observed data is consistent with moderate-to-large treatment effects in favor of the new therapy [3,4]. (b) As will be shown in the section 'Comparisons of the inefficacy monitoring rules,' a number of frequently used inefficacy rules are too conservative in the middle of the study and thus may allow the study to continue past the point where sufficient evidence to abandon the new therapy is available. (c) Many RCT monitoring plans do not commence monitoring early enough to maximize the possibility of early stopping in case the new therapy is harmful [4]. These considerations suggest that the commonly used interim monitoring methodologies may not be optimal for designing effective inefficacy monitoring rules. Furthermore, the continuing use of these methodologies suggests that their deficiencies may often not be fully appreciated by the investigators designing trials.

After discussing some general issues with inefficacy monitoring and defining some commonly used rules we propose a general inefficacy monitoring procedure that has a number of desirable properties: (a) it will stop the trial very early if the therapy is harmful, (b) it will stop the trial if the interim data provide convincing evidence that the new therapy is inefficacious, (c) the proposed stopping boundary will be easily understood, including by nonstatisticians, and (d) the use of the procedure will potentially result in fewer treated patients and earlier release of results under the null hypothesis with only a very minor loss of power under the alternative hypothesis. In the section 'Comparisons of the inefficacy monitoring rules,' we present the results of some limited simulations comparing the proposed procedure with existing procedures. We end with a discussion of the major considerations for constructing inefficacy monitoring guidelines.

Incorporating an inefficacy boundary into RCT design

A typical superiority RCT is designed to have adequate power (80–90%) using an appropriate type I error level (e.g., 0.025 one-sided) to detect the minimal clinically relevant benefit. To account for the multiple testing, efficacy monitoring boundaries adjust nominal significance levels for test statistics at interim and the final analyses to maintain the overall type I error level specified in the trial design. Under the null hypothesis, the addition of the possibility of inefficacy stopping reduces the overall type I error of the design. From a theoretical standpoint, it is therefore possible to 'reclaim' unspent type I error by adjusting the upper (efficacy) boundary downwards to bring the overall type I error up to the level specified in the design [5–7]. However, this formal approach is inconsistent with clinical trial practice because interim-monitoring boundaries are generally recognized to be guidelines rather than strict rules. This is especially relevant to inefficacy stopping decisions that are often based on a synthesis of benefit, safety and feasibility considerations both

internal and external to the trial (by contrast, efficacy stopping is usually driven by internal evidence). As a result, adjusting of efficacy boundary to reclaim the type I error reduced by inefficacy monitoring is not recommended [8–10]. In what follows, we will assume that the trial is designed so that the type I error is controlled in the absence of inefficacy monitoring.

Under the alternative hypothesis, the addition of the possibility of inefficacy stopping reduces the power of the trial design. In theory, one could increase the sample size to account for this. However, as the inefficacy boundaries we consider generally result in less than 1% loss of power, we will not increase the sample size to account for the inefficacy monitoring.

In practice, many RCTs use one of the following three general methods to define an inefficacy monitoring rule:

(1) *Group sequential testing of the alternative hypothesis*: The study is stopped if the alternative hypothesis is rejected [7,11]. To avoid substantial loss of power (inflation of type II error) due to multiple testing, the nominal significance levels at given interim-monitoring timepoints are determined according to a pre-specified spending function (e.g., O'Brien-Fleming type boundary [12]). Alternatively, a small constant nominal significance level is used. In what follows, we shall use a nominal significance level of 0.0025 as was used by Petrylak *et al.* [13], Williamson *et al.* [14], and Lara *et al.* [15].
(2) *Repeated confidence intervals (RCI) approach [16, p. 199]*: The study is stopped if a pre-specified level RCI excludes the alternative hypothesis at any of the designated interim-monitoring timepoints. In what follows, we shall consider repeated confidence intervals using an O'Brien-Fleming spending function with overall one-sided 0.025 level (one-sided 97.5% RCI) as was used by Stadtmauer *et al.* [17], Sandler *et al.* [18], Giantonio *et al.* [19], and Miller *et al.* [20].
(3) *Conditional power (stochastic curtailment) approach [2,21]*: The study is stopped if, given the observed data, the conditional power (that the trial would reject the null hypothesis if completed) under a specific assumption about the treatment effect in the future data is less than a pre-specified threshold (generally in the 10–30% range [2,22]). The conditional power calculation requires making an assumption about the distribution of the future data [23]; this can be considered as a desirable property [24], or not [25]. This approach is, by our nomenclature, futility monitoring. In what follows, we will consider a conditional power rule that is often used to prospectively define the inefficacy boundary: stop for inefficacy if the conditional power under the alternative hypothesis is less than 10%, as was used by Teerlink *et al.* [26], The TRIUMPH investigators [27], Spriggs *et al.* [28], and Rerks-Ngarm *et al.* [29]. Although somewhat uncommon as a prospectively defined rule, a conditional power approach with a 30% threshold is also included.

Another approach to inefficacy monitoring is with the use of predictive power [30,16]. It is sometimes used to aid DMC deliberations, for example, Dignam *et al.* [31]. This approach requires specifying a prior distribution for the treatment effect and is seldom used for a prospective definition of the inefficacy boundary. Therefore, we do not include predictive power in our comparisons.

A new approach to inefficacy monitoring

Our proposal will be presented for a time-to-event endpoint (survival setting), which is typical for RCTs in cancer or cardiovascular diseases (the method is naturally generalizable to other outcomes types). In this setting the observed treatment effect is usually summarized by the observed hazard ratio $\hat{\theta}$ (experimental arm over control arm). With this parameterization, observing $\hat{\theta}$ greater than one corresponds to the new therapy being worse than the control arm.

An intuitively attractive and simple inefficacy rule can be based on whether after one-half of the total study information is available the estimate of the treatment effect indicates that the new therapy is worse than the control treatment, that is, $\hat{\theta} > 1$ [32,33]. Rather than performing this monitoring arbitrarily at one-half the information, we suggest beginning the monitoring at the earliest timepoint at which $\hat{\theta} > 1$ would imply that the two-sided 95% confidence interval for log θ would not contain the log alternative hypothesis used to design the trial (log θ_A). Stopping the trial at this time would provide conventionally convincing evidence that the treatment is inefficacious (in the sense that the data are inconsistent with the design alternative) while at the same time being intuitively attractive (the results are going in the wrong direction). When expressed as a fraction of the total study information, this earliest point t_0 does not depend on the design alternative and is given by:

$$t_0 = \left(\frac{C_{0.975}}{C_{1-\alpha/2} + C_{1-\beta}} \right)^2 \quad (1)$$

for an RCT designed with two-sided significance level α and power $1-\beta$, where C_η is η quantile of the standard normal distribution; see Appendix A. For a one-sided 0.025 level designs, these information fractions are 37%, 43%, and 49% for 90%, 85%, and 80% power designs, respectively.

A comprehensive inefficacy monitoring rule should allow repeated examination of the accumulating data. With increasing information, the evidence contained in observed treatment effect and the corresponding 95% confidence interval may allow one to demonstrate inefficacy even when the experimental therapy appears to be doing (slightly) better than the control therapy. This suggests a simple rule where one stops the trial for inefficacy if at any point the 95% confidence interval for the log hazard ratio does not contain the design-alternative log hazard ratio. (This should not be confused with the repeated confidence interval approach described in the previous section – here we are using 95% confidence intervals at each look unadjusted for multiple comparisons.) A problem with this simple approach is that a trial can be stopped for inefficacy even though the experimental therapy is doing moderately better than the control treatment near the end of the trial. For example, with a trial designed with 90% power to detect a hazard ratio of 0.75, the 95% confidence interval for the hazard ratio would not cover the alternative even when the observed hazard ratio was 0.90 near the end of the trial. That is, the trial would stop early for inefficacy even though the observed hazard ratio was 0.90 favoring the experimental therapy. In many settings this may not be sufficient to convince the clinical community that the therapy has no tangible benefit. Similarly, this simple approach can be too aggressive very early in the trial: for example, at 25% of information the rule would correspond to stopping with an observed hazard ratio of 1.06 (one-sided p-value 0.67 against the null hypothesis) for a 90% power trial targeting a hazard ratio of 0.75.

To avoid the potential problem of stopping the trial for inefficacy when the experimental arm is doing moderately better, we propose the following. The inefficacy monitoring boundary starts with the cut-off $\hat{\theta} > 1$ at information time t_0 (given by Equation (1)) and gradually allows for the cut-off to *increase* to f% of the targeted benefit subject to the requirement that a two-sided 95% confidence interval for the treatment effect observed at a cut-off excludes the alternative hypothesis. A boundary (on a log hazard ratio scale) satisfying the above conditions can be constructed by taking f to be 20% and connecting the points $(t_0, 0)$ and $(1, 0.20 * \log\theta_A)$ by a straight line. Then for a given information fraction t in the interval $[t_0,1]$, the corresponding cut-off value is (see Appendix A for details):

$$0.20 * \log\theta_A \frac{\left(C_{1-\alpha/2} + C_{1-\beta}\right)^2 t - C_{0.975}^2}{\left(C_{1-\alpha/2} + C_{1-\beta}\right)^2 - C_{0.975}^2} \quad (2)$$

For example, with a trial designed with 90% power to detect a hazard ratio of 0.75, the boundary goes from 0 to −0.058 on the log hazard ratio scale, or from 1.00 to 0.94 on the hazard ratio scale. In practice the approach is implemented as follows: inefficacy monitoring is scheduled to start approximately after information fraction t_0 becomes available with repeated analyses allowed as frequently as at 10% information increment with the cut-off values given by Equation (2). We refer to this proposed boundary as the Linear 20% Inefficacy Boundary (LIB20). Choices other than 20%, for example, LIB40, etc., are discussed in the Discussion.

Although the proposed inefficacy rule works well in terms of stopping under the null hypothesis of no difference between the study arms ($\theta = 1$), additional stopping criteria are needed to protect against the situation in which the new treatment is nontrivially less efficacious than the control therapy. For example, a 461-patient RCT evaluating the experimental drug Canfosfamide in ovarian cancer reported 8.5- versus 13.6-month median overall survival on the experimental and the control treatment arms, respectively ($p = 0.0001$) [34]. To allow adequate protection for the study participants under such a scenario, inefficacy monitoring should also allow the possibility of stopping the trial as soon as possible if the experimental therapy appears clearly to be harmful. This can be accomplished by adding an early 'harm' look before the data are mature enough for inefficacy evaluation. We suggest performing the harm look at 25% information using the rule that stops for harm if the lower 95% (unadjusted) confidence bound for the hazard ratio is above 1. This is the same as one-sided 0.05 level test of the null hypothesis in the direction of harm. Therefore, our proposed approach is LIB20 with a harm look at 25% information. In addition, for the simulation study in the next section we modify the unadjusted 95% confidence interval approach to use the unadjusted 95% CI rule in the interval $[t_0,1]$ and a harm look at 25% information.

Comparisons of the inefficacy monitoring rules

We considered two scenarios for evaluation of the inefficacy rules. The first is based on three interim looks: for a 90% power trial, looks at 25%, 40%, and

70% of information, and for an 80% power trial, looks at 25%, 50%, and 75% of information. The second scenario employs a more frequent monitoring schedule: for a 90% power trial, a look at 25% information followed by six looks performed at 10% information increments starting at 40% of information, and for an 80% power trial, a look at 25% information followed by five looks performed at 10% information increments starting at 50% of information. We consider the monitoring boundaries implied by the different monitoring rules on two different scales: the normalized test statistic for testing the null hypothesis (Z-value scale) and the observed hazard ratio. These boundaries are presented in Tables 1 and 2, along with the value for the lower one-sided 97.5% confidence bound that would be observed if the data just crossed the corresponding inefficacy boundary for a trial targeting an alternative hazard ratio of 0.75; see Appendix A for details of the boundary construction. We also present in these tables the corresponding values at that final analysis (100% information) that would be used to decide if the null hypothesis was rejected (if the trial did not stop early).

We examine the operating characteristics of the monitoring rules below, but the monitoring boundaries themselves suggest some problems for some of the rules. As noted in the previous section the unadjusted 95% confidence interval rule can stop near the end of the trial even though the observed hazard ratio is 0.90 (Table 1). A similar problem is seen with the 10% conditional power rule and the repeated one-sided 97.5% confidence interval rule, where the observed hazard ratio can be 0.89 and 0.92, respectively. (One could argue that at the final analysis the rejection region is being determined by a cut-off of 0.84, so there is no problem here; we return to this point in the Discussion.) For the 30% conditional power rule the degree of aggressiveness in the second part of the trial makes it unsuitable in many clinical settings. For example, for an 80%-power design (Table 2) the rule suggests inefficacy stopping with the observed hazard ratio of 0.84 (for a design where 10% of information later observing a hazard ratio of 0.82 would be considered a positive result). At 25% of information, the proposed rule stops when the null hypothesis can be rejected in favor of superior efficacy of the control arm at the one-sided 0.05 level ($Z < -1.645$). We believe the 10% conditional power rule and the repeated confidence interval rule are too conservative here, requiring with seven looks, $Z < -3.18$ and $Z < -2.53$, respectively (Table 1). Finally, note that the lower one-sided 97.5% confidence interval can be below 0.75 at stopping for the conditional power rules (Tables 1 and 2). This is problematic because the trial could be stopped for inefficacy even though the data are not inconsistent with the alternative hypothesis.

We performed a limited simulation to investigate the operating characteristics of the inefficacy monitoring rules (see Appendix B for details). In particular, we compared the rules in terms of their loss of power, their probability of stopping under the null hypothesis, and their average stopping time under the null hypothesis. The average stopping time was calculated using three scales: information time (proportion of the maximum information), calendar time (proportion of the maximum trial duration), and sample size (proportion of the maximum sample size). The information at a stopping time was translated to a calendar time scale and an accrual scale assuming uniform accrual over 4 years, a follow-up period of 2 years, a hazard ratio of 0.75, and an exponential survival assumption with median survival of 2 years in the control treatment arm.

The simulation results are presented in Tables 3 and 4. All three commonly used approaches (with the exception of the 30% conditional power rule) have later average stopping times than the proposed approach. Of particular concern are nontrivially higher stopping times on both calendar time and patient accrual scales. For example, when the new therapy is ineffective, a 90%-power design using the repeated CI approach with three looks would on average continue to 80% of the total study duration and enroll 97% of the total required number of patients. In comparison, our approach would on average be stopped at 70% of the total study duration and enroll 88% of the total number of patients. The unadjusted 95% confidence interval rule has slightly better stopping characteristics than the proposed rule, but, in addition to the problems with the boundary mentioned above, it has considerable loss of power due to the inefficacy monitoring: 1.3% and 1.8% absolute loss for 90% power trials with three and seven interim looks, respectively. The 30% conditional power rule is inferior to the proposed approach for a 90%-power design. While the 30% conditional power rule has better average stopping times for an 80%-power design, it results in an unacceptable 3.3% loss of power.

The overall probability of stopping under H_0 (third column in Tables 3 and 4) should be interpreted with caution: it does not account for when the trial is stopped, and thus may not reflect the true patient/resource impact of stopping. For example, in Table 3 for 7-look rules the 10% conditional power approach has a higher overall probability of stopping than the proposed approach (91% vs. 79%, respectively). However, the 10% conditional power boundary tends to stop later into the trial and thus is considerably less efficient in terms of

Table 1 Inefficacy stopping boundaries 90% power (for a trial targeting hazard ratio of 0.75)

Monitoring plan	Boundary values including the final analysis															
	Z-value scale Hazard-ratio scale[a]								Lower one-sided 97.5% confidence bound (unadjusted)							
	3 interim looks															
	25%	40%			70%			100%	25%	40%			70%			100%
Proposed (LIB20 with harm look)	−1.645 1.34	0.02 1.0			0.29 0.97			1.96 0.84	0.95	0.76			0.79			0.707
Unadjusted two-sided 95% CI (+ harm look)	−1.645 1.34	0.1 0.99			0.76 0.92			1.96 0.84	0.95	0.75			0.75			0.707
Repeated CI rule (97.5% one-sided)	−2.37 1.52	−1.11 1.17			0.33 0.97			1.96 0.84	1.08	0.89			0.78			0.707
Testing H_a rule (0.0025 one-sided)	−1.18 1.23	−0.75 1.11			−0.08 1.01			1.96 0.84	0.87	0.84			0.82			0.707
Conditional power (<10%)	−3.18 1.76	−1.56 1.24			0.34 0.97			1.96 0.84	1.24	0.95			0.78			0.707
Conditional power (<30%)	−1.87 1.39	−0.63 1.09			0.83 0.92			1.96 0.84	0.98	0.83			0.74			0.707
	7 interim looks															
	25%	40%	50%	60%	70%	80%	90%	100%	25%	40%	50%	60%	70%	80%	90%	100%
Proposed (LIB20 with harm look)	−1.645 1.34	0.02 1.0	0.10 0.99	0.19 0.98	0.29 0.97	0.40 0.96	0.52 0.95	1.96 0.84	0.95	0.76	0.77	0.78	0.79	0.79	0.79	0.707
Unadjusted two-sided 95%CI (+ harm look)	−1.645 1.34	0.1 0.99	0.34 0.96	0.56 0.94	0.76 0.92	0.95 0.91	1.13 0.90	1.96 0.84	0.95	0.75	0.75	0.75	0.75	0.75	0.75	0.707
Repeated CI rule (97.5% one-sided)	−2.53 1.57	−1.23 1.19	−0.64 1.08	−0.17 1.02	0.24 0.98	0.58 0.94	0.89 0.92	1.96 0.84	1.11	0.90	0.85	0.81	0.79	0.78	0.77	0.707
Testing H_a rule (0.0025 one-sided)	−1.18 1.23	−0.75 1.11	−0.51 1.07	−0.29 1.03	−0.08 1.01	0.10 0.99	0.28 0.97	1.96 0.84	0.87	0.84	0.83	0.83	0.82	0.82	0.81	0.707
Conditional power (<10%)	−3.18 1.76	−1.56 1.24	−0.81 1.11	−0.20 1.02	0.34 0.97	0.82 0.92	1.30 0.89	1.96 0.84	1.24	0.95	0.87	0.82	0.78	0.76	0.74	0.707
Conditional power (<30%)	−1.87 1.39	−0.63 1.09	−0.05 1.01	0.42 0.95	0.83 0.92	1.20 0.89	1.55 0.87	1.96 0.84	0.98	0.83	0.79	0.76	0.74	0.73	0.72	0.707

[a]Hazard ratios less than one mean the experimental treatment arm is doing better than the control treatment arm.

Table 2 Inefficacy stopping boundaries 80% power (for a trial targeting hazard ratio of 0.75)

Monitoring plan	Boundary values including the final analysis													
	Z-value scale Hazard-ratio scale[a]							Lower 97.5% confidence bound (unadjusted)						
	3 interim looks													
	25%	50%		75%			100%	25%	50%		75%			100%
Proposed (LIB20 with harm look)	−1.645 1.4	0.01 1.0		0.25 0.97			1.96 0.82	0.94	0.75		0.77			0.670
Unadjusted two-sided 95% CI (+ harm look)	−1.645 1.4	0.03 1.0		0.48 0.94			1.96 0.82	0.94	0.75		0.75			0.670
Repeated CI rule (97.5% one-sided)	−2.62 1.71	−0.86 1.13		0.12 0.99			1.96 0.82	1.14	0.85		0.78			0.670
Testing H_a rule (0.0025 one-sided)	−1.40 1.33	−0.81 1.12		−0.37 1.04			1.96 0.82	0.89	0.85		0.83			0.670
Conditional power (< 10%)	−2.53 1.68	−0.50 1.08		0.71 0.92			1.96 0.82	1.12	0.81		0.73			0.670
Conditional power (< 30%)	−1.22 1.28	0.25 0.96		1.15 0.87			1.96 0.82	0.86	0.73		0.69			0.670
	6 interim looks													
	25%	50%	60%	70%	80%	90%	100%	25%	50%	60%	70%	80%	90%	100%
Proposed (LIB20 with harm look)	−1.645 1.4	0.01 1.0	0.09 0.99	0.19 0.98	0.31 0.97	0.43 0.95	1.96 0.82	0.94	0.75	0.76	0.77	0.77	0.77	0.670
Unadjusted two-sided 95% CI (+ harm look)	−1.645 1.4	0.03 1.0	0.22 0.97	0.40 0.95	0.56 0.94	0.71 0.93	1.96 0.82	0.94	0.75	0.75	0.75	0.75	0.75	0.670
Repeated CI rule (97.5% one-sided)	−2.75 1.75	−0.95 1.15	−0.50 1.07	−0.13 1.02	0.19 0.98	0.48 0.95	1.96 0.82	1.18	0.86	0.83	0.80	0.78	0.77	0.670
Testing H_a rule (0.0025 one-sided)	−1.40 1.33	−0.81 1.12	−0.62 1.09	−0.45 1.06	−0.29 1.03	−0.13 1.01	1.96 0.82	0.89	0.85	0.84	0.83	0.83	0.82	0.670
Conditional power (< 10%)	−2.53 1.68	−0.50 1.08	0.03 1.0	0.49 0.94	0.92 0.90	1.34 0.87	1.96 0.82	1.12	0.81	0.77	0.74	0.72	0.70	0.670
Conditional power (< 30%)	−1.22 1.28	0.25 0.96	0.65 0.92	0.99 0.89	1.30 0.86	1.59 0.84	1.96 0.82	0.86	0.73	0.71	0.70	0.69	0.68	0.670

[a]Hazard ratios less than one mean the experimental treatment arm is doing better than the control treatment arm.

Table 3 Trials designed with 90% power (0.025 one-sided significance level, accrued over 4 years with 2 years of follow-up)

Boundary	Loss of power[a] due to inefficacy monitoring under H_A (%)	Probability of inefficacy stopping under H_0 (%)	Average stopping time under H_0		
			Information scale (%)	Calendar time scale (%)	Sample size scale (%)
Monitoring plan with 3 interim looks (25%, 40%, and 70% of information)					
Proposed (LIB20 with harm look)	0.8	68	64	70	88
Unadjusted 95% CI (with harm look)	1.3	80	59	66	87
Repeated CI rule (97.5% one-sided)	<0.1	63	77	80	97
Testing H_a rule (0.0025 one-sided)	0.1	50	76	80	92
Conditional power rule ($<10\%$)	<0.1	63	79	81	99
Conditional power rule ($<30\%$)	0.5	80	68	72	93
Monitoring plan with 7 interim looks (25%, 40%, 50%, 60%, 70%, 80%, 90% of information)					
Proposed (LIB20 with harm look)	1.0	79	59	66	86
Unadjusted 95% CI (with harm look)	1.8	91	54	62	85
Repeated CI rule (97.5% one-sided)	0.1	83	70	74	95
Testing H_a rule (.0025 one-sided)	0.1	67	70	75	91
Conditional power rule ($<10\%$)	0.2	91	70	73	97
Conditional power rule ($<30\%$)	1.4	95	59	65	91

[a]Absolute loss of power.

Table 4 Trials designed with 80% power (.025 one-sided significance level, accrued over 4 years with 2 years of follow-up)

Boundary	Loss of power[a] due to futility monitoring under H_A (%)	Probability of inefficacy stopping under H_0 (%)	Average stopping time under H_0		
			Information scale (%)	Calendar time scale (%)	Sample size scale (%)
Monitoring plan with 3 interim looks (25%, 40%, and 75% of information)					
Proposed (LIB20 with harm look)	0.3	66	70	75	93
Unadjusted 95%CI (with harm look)	0.4	72	68	73	93
Repeated CI rule (97.5% one-sided)	<0.1	55	81	83	98
Testing H_a rule (0.0025 one-sided)	0.1	40	82	85	95
Conditional power rule ($<10\%$)	0.2	76	73	76	96
Conditional power rule ($<30\%$)	1.9	88	60	67	90
Monitoring plan with 6 interim looks (25%, 50%, 60%, 70%, 80%, 90% of information)					
Proposed (LIB20 with harm look)	0.4	75	67	72	93
Unadjusted 95% CI (with harm look)	0.5	82	65	70	92
Repeated CI rule (97.5% one-sided)	<0.1	71	77	80	98
Testing H_a rule (0.0025 one-sided)	0.1	52	79	82	95
Conditional power rule ($<10\%$)	0.5	91	67	72	96
Conditional power rule ($<30\%$)	3.3	95	56	64	90

[a]Absolute loss of power.

the average stopping times, for example, average accrual is 97% versus 86%, respectively.

The operating characteristics in Tables 3 and 4 do not account for any interim monitoring for efficacy that might be used. Monitoring for efficacy would have very little effect on the operating characteristics under H_0, but can reduce the loss of power due to the inefficacy monitoring under H_A. For example, using the O'Brien-Fleming boundary for efficacy, the loss of power due to inefficacy monitoring using the proposed rule in a 90% power trial would be 0.8% and 0.4% with three and seven looks, respectively. Using a Haybittle-Peto [35] efficacy monitoring rule (i.e., stop for efficacy if the one-sided $p < 0.001$), the corresponding figures are 0.8% and 0.9%. These figures can be compared to the losses of

power due to inefficacy monitoring when there is no efficacy monitoring of 0.8% and 1.0% (column 2, Table 3). Intuitively, the addition of efficacy monitoring to a clinical trial design raises the bar for rejecting the null hypothesis at the final analysis, omitting many of the borderline positive trials that would have resulted in lost power due to the inefficacy monitoring. (The additional trials that would be stopped early for efficacy and contribute to the power would unlikely have been stopped for inefficacy.)

Discussion

Not rejecting the null hypothesis at the final analysis is qualitatively different from stopping a trial early for inefficacy. In the latter situation, the data should clearly imply that the experimental treatment is not useful as a replacement for the control treatment. In the former situation, although the experimental treatment is disappointing, it still may have clinical utility in some settings. For example, consider a relatively nontoxic experimental treatment that just fails to reject the null hypothesis. The two-sided confidence interval for the treatment effect will cover the design alternative, and the observed treatment effect may appear clinically interesting. For example, the lower 97.5% confidence bound for the treatment effect would be 0.707 and 0.670 for trials designed with 90% and 80% power for the alternative hazard ratio 0.75 (Tables 1 and 2).

Because of the distinction between stopping early and rejecting the null hypothesis at the final analysis, we have suggested that a requirement for early stopping for inefficacy is that the two-sided 95% confidence interval for the treatment effect should not cover the alternative hypothesis. This requirement still leaves many possibilities for an interim monitoring boundary. In addition to a harm look at 25% of information, we have suggested starting the inefficacy monitoring at 40%, 45%, and 50% of information for 90%, 85%, and 80% power trial designs, respectively. At these time points, the boundary implied by the 95% confidence interval not containing the alternative hypothesis is approximately equal to the observed hazard ratio being greater than 1, an appealing intuitive rule.

The proposed inefficacy approach can be easily adjusted to accommodate specific challenges of a given clinical setting. In some settings, the harm look at 25% of information may be considered premature, especially when there is concern about crossing hazards [36]. In such cases, it may be appropriate to perform this look later or omit it completely. Even without crossing hazards, the hazards may not be proportional when the effect of the experimental treatment is delayed. This situation might be more prevalent with treatments for progressive diseases such as diabetes and coronary artery disease, as opposed to diseases such as cancer or heart failure. In the settings with delayed treatment effects, a more conservative approach to inefficacy monitoring than being proposed here may be appropriate, for example, by beginning inefficacy monitoring at a later time. However, note that a major delayed treatment effect leading to serious nonproportional hazards calls into question a definitive analysis based on proportional hazards. With a definitive analysis not based on proportional hazards (e.g., a nonproportional hazards cure model), the inefficacy monitoring will need to be tailored to the definitive analysis plan.

A different clinical setting is when the new therapy is relatively toxic (or when the control arm contains no active therapy). In this setting, the proposed approach may be adjusted to provide a more aggressive inefficacy monitoring in the later parts of the study by using LIB40 instead of LIB20. Another possibility is to use the rule implied by the unadjusted 95% confidence intervals not containing the target alternative (with a harm look at 25% of information). On the other hand, if LIB20 boundary was felt to be too aggressive, then one might use the more conservative boundary LIB0, which corresponds to repeatedly examining the data to see if the hazard ratio is greater than 1.

Our inefficacy interim monitoring approach is attractive from statistical, clinical, and logistical standpoints. By decreasing average stopping times relative to the commonly used boundaries, our rule lessens patient exposure to inactive treatments, improves resource utilization, and accelerates dissemination of important clinical information. At the same time, the proposed rule provides a clear benchmark for providing compelling evidence that the new therapy is not beneficial that can be easily understood by members of multidisciplinary data monitoring committees and appreciated by both clinicians and patients. The proposed inefficacy boundary does not depend on the study design parameters (targeted effect or sample size), and is based on applying a simple decision rule allowing a flexible analysis schedule with trivial power loss ($< 1\%$). Therefore, when appropriate, it can be easily added to an existing RCT design without requiring any sample size adjustment.

Acknowledgement

We would like to thank the referees and the editor for their helpful comments.

References

1. **Ellenberg SS, Fleming TR, DeMets DL.** *Data Monitoring Committees in Clinical trials*. Willey, Chichester, 2002.
2. **Ware JH, Muller JE, Braunwald E.** The futility index. *Am J Med* 1985; **78**: 635–43.
3. **Freidlin B, Korn EL.** A comment on futility monitoring. *Control Clin Trials* 2002; **23**: 355–66.
4. **Freidlin B, Korn EL.** Monitoring for lack of benefit: a critical component of a randomized clinical trial. *J Clin Oncol* 2009; **27**: 629–33.
5. **Whitehead J, Stratton I.** Group sequential clinical trials with triangular continuation regions. *Biometrics* 1983; **39**: 227–36.
6. **Emerson SS, Fleming TR.** Symmetric group sequential test designs. *Biometrics* 1989; **45**: 905–23.
7. **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. *J Stat Plan Inf* 1994; **42**: 19–35.
8. **European Medicines Agency.** Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. Available at: http://www.emea.europa.eu/pdfs/human/ewp/245902enadopted.pdf (accessed 26 March 2010).
9. **Snapinn S, Chen MG, Jiang Q, *et al*.** Assessment of futility in clinical trials. *Pharm Stat* 2006; **5**: 273–81.
10. **Cook TD, DeMets DL.** *Introduction to Statistical Methods for Clinical Trials*. Chapman & Hall/CRC, Boca Raton, 2008.
11. **Pocock SJ.** When to stop a clinical trial. *BMJ* 1992; **305**: 235–40.
12. **O'Brien PC, Fleming TR.** A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**: 549–56.
13. **Petrylak DP, Tangen CM, Hussain MH, *et al*.** Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *N Engl J Med* 2004; **351**: 1513–20.
14. **Williamson SK, Crowley JJ, Lara PN, *et al*.** Phase III trial of paclitaxel plus carboplatin with or without tirapazamine in advanced non-small-cell lung cancer: Southwest Oncology Group Trial S0003. *J Clin Oncol* 2005; **23**: 9097–104.
15. **Lara Jr PN, Natale R, Crowley J, *et al*.** Phase III trial of irinotecan/cisplatin compared with etoposide/cisplatin in extensive-stage small-cell lung cancer: clinical and pharmacogenomic results from SWOG S0124. *J Clin Oncol* 2009; **27**: 2530–35.
16. **Jennison C, Turnbull BW.** *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton, 2000.
17. **Stadtmauer EA, O'Neill A, Goldstein LJ, *et al*.** Conventional-dose chemotherapy compared with high-dose chemotherapy plus autologous hematopoietic stem-cell transplantation for metastatic breast cancer. Philadelphia Bone Marrow Transplant Group. *N Engl J Med* 2000; **342**: 1069–76.
18. **Sandler A, Gray R, Perry MC, *et al*.** Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N Engl J Med* 2006; **355**: 2542–50.
19. **Giantonio BJ, Catalano PJ, Meropol NJ, *et al*.** Bevacizumab in combination with oxaliplatin, fluorouracil, and leucovorin (FOLFOX4) for previously treated metastatic colorectal cancer: results from the Eastern Cooperative Oncology Group Study E3200. *J Clin Oncol* 2007; **25**: 1539–44.
20. **Miller K, Wang M, Gralow J, *et al*.** Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. *N Engl J Med* 2007; **357**: 2666–76.
21. **Lan KKG, Simon R, Halperin M.** Stochastically curtailed tests in long-term clinical trials. *Commu Statist C* 1982; **1**: 207–19.
22. **Demets DL.** Futility approaches to interim monitoring by data monitoring committees. *Clin Trials* 2006; **3**: 522–29.
23. **Lan KKG, Wittes J.** The B-value: a tool for monitoring data. *Biometrics* 1988; **44**: 579–85.
24. **Evans SR, Li L, Wei LJ.** Data monitoring in clinical trials using prediction. *Drug Inform J* 2007; **41**: 733–42.
25. **Lachin JM.** Futility interim monitoring with control of type I and II error probabilities using the interim Z-value or confidence limit. *Clin Trials* 2009; **6**: 565–73.
26. **Teerlink JR, McMurray JJ, Bourge RC, *et al*.** Tezosentan in patients with acute heart failure: design of the Value of Endothelin Receptor Inhibition with Tezosentan in Acute heart failure Study (VERITAS). *Am Heart J* 2005; **150**: 46–53.
27. **The TRIUMPH Investigators**. Effect of tilarginine acetate in patients with acute myocardial infarction and cardiogenic shock: the TRIUMPH randomized controlled trial. *JAMA* 2007; **297**: 1657–66.
28. **Spriggs DR, Brady MF, Vaccarello L, *et al*.** Phase III randomized trial of intravenous cisplatin plus a 24- or 96-hour infusion of paclitaxel in epithelial ovarian cancer: a Gynecologic Oncology Group Study. *J Clin Oncol* 2007; **25**: 4466–71.
29. **Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, *et al*.** Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N Engl J Med* 2009; **361**: 2209–20.
30. **Spiegelhalter DJ, Freedman LS, Blackburn PR.** Monitoring clinical trials: conditional or predictive power? *Control Clin Trials* 1986; **7**: 8–17.
31. **Dignam JJ, Bryant J, Wieand HS, *et al*.** Early stopping of a clinical trial when there is evidence of no treatment benefit: protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. *Control Clin Trials* 1998; **19**: 575–88.
32. **Ellenberg SS, Eisenberger MA.** An efficient design for phase III studies of combination chemotherapies. *Cancer Treat Rep* 1985; **69**: 1147–52.
33. **Wieand S, Schroeder G, O'Fallon JR.** Stopping when the experimental regimen does not appear to help. *Stat Med* 1994; **13**: 1453–58.
34. **Vergote I, Finkler N, Del Campo J, *et al*.** Single agent canfosfamide vs pegylated liposomal doxorubicin or topotecan 3rd-line treatment of platinum refractory or resitant ovarian cancer. 2007; Annual Meeting of the American Society of Clinical Oncology, Abstract LBA5528.
35. **Haybittle JL.** Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 1971; **44**: 793–97.
36. **Korn EL, Freidlin B, Mooney M.** Stopping or reporting early for positive results in randomized clinical trials: the NCI Cooperative Group Experience 1990–2005. *J Clin Oncol* 2009; **27**: 1543–44.

Appendix A: Linear inefficacy boundary

Background

Consider a survival setting under a proportional hazards model where the ratio of the experimental over the control arm hazard is denoted by θ. It can be shown (see, Jennison and Turnbull [16], for example) that for an interim analysis done at the

information fraction t and the final analysis done at the full information I the joint distribution of the corresponding standardized log-rank statistics Z_t and Z is approximately bivariate normal with mean $(\log\theta\sqrt{tI},\ \log\theta\sqrt{I})$, and covariance $\sqrt{t}$.

For a trial designed to detect the hazard ratio θ_A with a given power $(1-\beta)$ at a two-sided significance level α the full (total) information is obtained from the equation:

$$\sqrt{I}\log\theta_A = C \tag{A.1}$$

where, for a fixed sample size trial, the constant $C=-(C_{1-\alpha/2}+C_{1-\beta})$ with C_η denoting the η quantile of a standard normal distribution.

The earliest point at which $\log\hat{\theta}>0$ implies that CI does not contain $\log\theta_A$

At the time point when the fraction t of the total information is available the width of a $(1-\gamma)$%CI for the log hazard ratio is $2*C_{1-\gamma/2}/\sqrt{tI}$. From (A.1), the earliest information fraction t_0 at which a $(1-\gamma)$%CI for log hazard ratio centered at 0 does not contain the design alternative is $\left(C_{1-\gamma/2}/(C_{1-\alpha/2}+C_{1-\beta})\right)^2$.

Construction of the linear inefficacy boundary

Consider a two-dimensional space with information fraction and inefficacy cut-off (on the log hazard ratio scale) as coordinates. This boundary is based on a straight line connecting:

(1) point $\left(\left(C_{1-\gamma/2}/(C_{1-\alpha/2}+C_{1-\beta})\right)^2,\ 0\right)$ corresponding to the cut-off of log(hazard-ratio) = 0 at the lowest information fraction where $(1-\gamma)$%CI excludes the alternative, and
(2) point $(1,\ f*\log\theta_A)$ corresponding to the f% fraction of the alternative at 100% information.

Then, for a RCT designed with two-sided significance level α and power $1-\beta$ at information fraction t in the interval $\left[\left(C_{1-\gamma/2}/(C_{1-\alpha/2}+C_{1-\beta})\right)^2,1\right]$ the boundary cut-off (on the log hazard-ratio scale) is:

$$f*\log\theta_A\frac{\left(\left(C_{1-\alpha/2}+C_{1-\beta}\right)^2 t-C_{1-\gamma/2}^2\right)}{\left(C_{1-\alpha/2}+C_{1-\beta}\right)^2-C_{1-\gamma/2}^2} \tag{A.2}$$

For $f\geq 0$ the proposed boundary has a number of desirable properties: it is monotone nondecreasing on both log-hazard ratio and Z scales.

Loss of power from the inefficacy monitoring does not depend on θ_A

We now show that for a RCT with a given power and significance level the loss of power from applying the proposed rule does not depend on the alternative hypothesis used in the design.

Without inefficacy analysis the power of a fixed sample size trial is $P_{\theta_A}\left(Z<-C_{1-\alpha/2}\right)=1-\beta$. Suppose an inefficacy look at information fraction t in $[t_0,1]$ using the cut-off given in formula (A.2) is added to the design. To simplify the presentation, formula (A.2) can be rewritten as $g(t,\alpha,\beta,\gamma)^*\log\theta_A$ where $g(t,\alpha,\beta,\gamma)$ is a function that does not depend on θ_A for a given α, β, and γ. Then with the hazard ratio estimated by $\hat{\theta}_t=\exp(Z_t/\sqrt{tI})$ the reduction in power is:

$$\begin{aligned}
&P_{\theta_A}\left(Z<-C_{1-\alpha/2}\right)\\
&\quad-P_{\theta_A}\left(\log\hat{\theta}_t<g(t,\alpha,\beta,\gamma)\log\theta_A, Z<-C_{1-\alpha/2}\right)\\
&=P_{\theta_A}\left(Z<-C_{1-\alpha/2}\right)\\
&\quad-P_{\theta_A}\left(Z_t/\sqrt{tI}<g(t,\alpha,\beta,\gamma)\log\theta_A, Z<-C_{1-\alpha/2}\right)\\
&=\Phi\left(-C_{1-\alpha/2}-\sqrt{I}\log\theta_A\right)\\
&\quad-\Phi_{2,\sqrt{t}}\left(\sqrt{tI}\log\theta_A(g(t,\alpha,\beta,\gamma)-1),\right.\\
&\quad\left.-C_{1-\alpha/2}-\sqrt{I}\log\theta_A\right)\\
&=\Phi\left(C_\beta\right)-\Phi_{2,\sqrt{t}}\left(\left(C_{1-\alpha/2}+C_\beta\right)\sqrt{t}(g(t,\alpha,\beta,\gamma)-1), C_\beta\right)
\end{aligned} \tag{A.3}$$

where the last equality is due to (A.1), Φ denotes the cumulative distribution function of the standard normal distribution and $\Phi_{2,\rho}$ denotes the cumulative distribution function of the standard bivariate normal distribution with correlation ρ. From the last line in (A.3) for a given α, β, γ, and t, the loss of power does not depend on θ_A.

For a trial with an efficacy boundary, in addition to $C_{1-\alpha/2}$ and $C_{1-\beta}$, the constant C in Equation (A.1) will also depend on the spending function that defines the boundary. However, C does not depend of θ_A. Therefore, similar calculations can be used to show that the above result applies to studies with efficacy boundaries. With multiple inefficacy looks the result is obtained in a similar manner.

Appendix B: Description of the simulation study

The simulations (500,000 replications) were conducted as follows. In each trial replication the

observed hazard ratios at each analysis time were generated using a normal approximation to the distribution of the log hazard ratio. The analyses times were defined on the information scale and the corresponding cut-offs for the stopping boundaries were approximated using a normal approximation to the log hazard ratio. It should be noted that in the survival setting, analyses times are usually based on the fraction of the total number of events. Therefore, our results that are based on the information fraction provide an approximation to the boundaries used in practice.