Clinical Trials 2006; **3**: 522–529 **ARTICLE** CLINICAL TRIALS

Futility approaches to interim monitoring by data monitoring committees

David L DeMets

Clinical trials involving participants with high risk of serious events or being exposed to a new intervention with potential serious risk are typically monitored during the course of the trial. Often, this monitoring activity is conducted by an independent group of experts, often referred to as a Data Monitoring Committee (DMC). The DMC responsibility includes monitoring for early evidence of a harmful effect or convincing evidence of a benefit. If the data are consistent and overwhelming, the DMC may recommend that a trial be terminated early. Trials may also be terminated early if the initial goal of demonstrating a benefit cannot in all likelihood be attained, or is futile. These are complicated issues and should be sorted out in the protocol and the DMC charter prior to the start of the trial before data begin to accumulate. Several statistical methods have been developed to assist a DMC in determining when a negative or harmful trend is substantial enough to render the trial continuation futile. These will be briefly summarized. However useful these methods might be, they alone are not adequate to make the decision and the DMC must take into account other issues to make a judgment. *Clinical Trials* 2006; **3**: 522–529. http://ctj.sagepub.com

Introduction

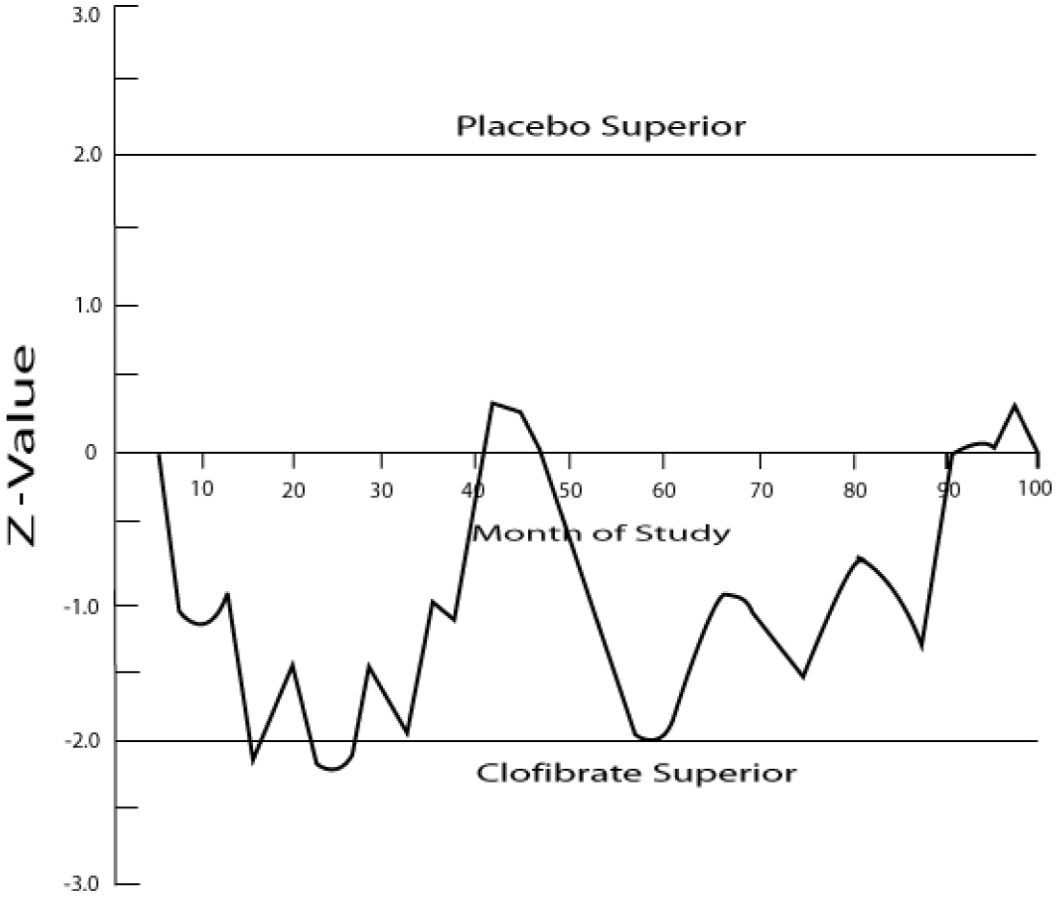
Monitoring of accumulating data in ongoing clinical trials is now common place and even required for trials with high risk patients or serious irreversible outcomes [1–3]. For many of these trials, the use of an independent panel of experts, referred to as a Data Monitoring Committee (DMC), is recommended [4]. The need and use for such committees was suggested in the late 1960s by a task force report, known as the Greenberg Report, commissioned by the National Institutes of Health (NIH) to describe how a multicenter randomized clinical trial should be organized and conducted [5]. While the frequency of use of DMCs increased during the next three decades, recent concerns about patient safety in clinical trials has only added to the interest in DMCs [6,7].

DMCs generally consist of three to five experts representing multiple disciplines such as clinical expertise, biostatistics, epidemiology, basic science and ethics. This group is charged with monitoring an ongoing trial for early and convincing evidence of a clinically important benefit or early evidence of harm. DMCs can recommend that a trial continue as planned, be modified or be terminated early. DMCs typically meet once or twice per year, depending on the rate of recruitment and rate of accumulating evidence, or when a fixed proportion of new information has accumulated (eg, every 20% increment).

Repeated examination of accumulating data can lead to false positive claims of benefit or harm unless special statistical procedures are utilized [8]. The false claim could be either claiming benefit or harm. Figure 1 summarizes the accumulating data from the Coronary Drug Project (CDP) which

University of Wisconsin-Madison, Department of Biostatistics and Medical Informatics, Madison, Wisconsin, USA
Author for Correspondence: David L. DeMets, University of Wisconsin-Madison, Department of Biostatistics and Medical Informatics, 600 Highland Avenue, Box 4675, Madison, WI 53792-4675, USA. E-mail: demets@ biostat.wisc.edu
Based on a presentation at the *10th International Symposium on Long-Term Clinical Trials*, Keble College, Oxford, UK, September 2005.

 10.1177/1740774506073115



Reprinted from *CONTROLLED CLINICAL TRIALS now known as CONTEMPORARY CLINICAL TRIALS* 1981; **1**: 363–76, with permission from Elsevier [9].

Figure 1 *Z*-values for clofibrate-placebo differences in proportion of deaths by calendar month since beginning of study (month 0 = March 1966, month 100 = July 1974)

compared survival for a cholesterol lowering drug, chlofibrate, with a placebo control [9]. The standardized statistic is plotted for the accumulating data with time in the study.

The conventional criteria for a nominal two-sided *P*-value 0.05 are indicated by the two parallel lines at ± 2.0. Of importance for this discussion is the substantial variability, especially early in the trial, which ultimately converges to a test statistic near zero, indicating no survival difference. Yet on several occasions, the nominal statistical criteria might have falsely suggested a benefit. Statistical methods have been developed which control the false positive error, also known as the Type I error, claiming a treatment effect when in fact none exists. These methods have been discussed extensively and summarized elsewhere [4,10,11].

Figure 2 depicts one commonly used criteria for establishing either benefit or harm, using group sequential boundaries as described by O'Brien and Fleming [12]. Upper and lower boundaries for a standardized test statistic are plotted on the horizontal axis versus the portion of the trial that has been completed, referred to as information fraction. The information fraction could be the fraction of planned patients recruited or the proportion of expected events observed. In this figure, a standardized statistic greater than the upper boundary would be considered consistent with evidence for benefit. A standardized statistic less that the lower negative boundary would be consistent with a treatment being harmful. (The boundary in between the upper and lower will be described below.)

In addition to terminating a trial early for evidence of treatment benefit or harm, the Greenberg Report recommended trials be terminated early if they could not meet their goals of establishing a beneficial treatment effect [5]. That is, if continuing the trial would be futile relative to establishing a treatment benefit, then perhaps the trial should be

terminated for ethical reasons. This might be a consideration if the interim analysis, comparing the experimental or new intervention with the control intervention, indicates a trend that is in the wrong or negative direction (a positive trend being consistent with benefit). However, an early negative trend might lead to a claim of no intervention effect due to chance alone or a small number of events, an error known as a Type II error. One goal in any trial is to minimize both Type I and II errors.

Clinical trials are usually not designed or conducted to establish that an intervention is harmful. However, there are circumstances that indicate the trial may need to distinguish whether an intervention is actually harmful or neutral, having ruled out all beneficial effects of clinical interest or importance [13]. Trials which are designed to rule out minimal intervention benefits are also called non-inferiority trials. Statistical methods have also been developed to aid in determining when a trial becomes futile. This analysis is often referred to as futility analysis, the topic of this paper.

Futility versus harm

There are trials where futility of continuing to the scheduled termination is an issue that the DMC must deal with but often this is not an easy task. A few examples illustrate this challenge.

The CONSENSUS-II trial was terminated early because the interim negative trend was no longer consistent with a beneficial effect on mortality and morbidity in a population of chronic heart failure patients [14]. The DMC decided that it was no longer ethical to continue and recommended early termination. The trial was designed to test if a drug, enalapril, delivered in a bolus to patients with acute myocardial infarction compared to a placebo control would reduce mortality. All patients received best standard of care as background therapy. A six-month comparison demonstrated a 10.2% mortality for placebo treated patients and 11% for the enalapril bolus treated patient, essentially ruling out a beneficial effect. In addition, a higher death rate of progressive heart failure in the enalapril group was nearing nominal significance. Since the bolus administration of this drug was not the standard, the data monitoring committee decided that there was no further benefit to patients being in the study and recommended termination.

The International Study of Infarct Survival (ISIS) provides another type of example [15]. In ISIS, atenolol was evaluated compared to a placebo control in approximately 16 000 patients with a myocardial infarction with mortality as the primary outcome variable. Early in this trial, a negative trend emerged near the nominal 0.05 significance. However, the data monitoring committee did not recommend termination since the number of events was small and to take into account the repeated testing effect. This negative trend continued for much of the trial. However, at the end of the planned follow-up, the negative trend reversed and became positive, showing a significant 15% reduction in mortality. Thus, early results although discouraging were not predictive of the ultimate results. Stopping this trial early for negative trends would have been a mistake, a false negative claim or a Type II error. Negative trends don't often reverse themselves to become beneficial and statistically significant but such cases do occur.

Another interesting circumstance may exist if an intervention with a negative trend for a primary outcome has other beneficial clinically relevant effects, such as reduction in symptoms or modification of other clinical outcomes. For example, recent examples of a class of drugs labeled as inotropic drugs were known in chronic heart failure patients to improve cardiac function and in some cases to improve quality of life. A series of trials evaluated such drugs for their ability to reduce mortality and also hospitalizations. In two of those trials, PROMISE and VEST, negative trends [16–18] emerged. At some point during these trials, the DMC could have become rather convinced that the final results were unlikely to prove beneficial, using methods described in the next section. However, even if these drugs were not going to be beneficial with respect to reducing mortality or hospitalization, they might still be of interest as long as they were not harmful. That is, these drugs improved patients' cardiac function and quality of life. Thus, distinguishing whether the interim negative trend would cross some lower boundary for harm or just be consistent with a lack of benefit would be important. If shown to be harmful, they would not likely be used or even approved for heart failure patients, despite some other beneficial effects. The interim results for mortality using the log rank test in the VEST are shown in Table 1. As shown, a negative trend emerged but with considerable variation in the early portion. Not until later in the trial, after a substantial number

Table 1 Accumulating results for VEST*

Information fraction	Log rank Z-value (high dose)
0.43	+0.99
0.19	−0.25
0.34	−0.23
0.50	−2.04
0.60	−2.32
0.67	−2.50
0.84	−2.22
0.90	−2.43
0.95	−2.71
1.0	−2.41

*VEST = Vesnarinone Trial [18].

of events had occurred, did the trends become stable. However, a previous trial of Vesnarinone had demonstrated a 60% reduction in total mortality ($P < 0.002$), contrary to the VEST trend [17].

Other examples of this type of negative trend are provided by recent trials of hormone replacement therapy (HRT). HRT became a widely used treatment for symptoms and the reduction of bone loss in postmenopausal women. Several epidemiologic studies also suggested that HRT would reduce cardiovascular disease in postmenopausal women, and was being utilized for that reason as well, despite no randomized trials showing a benefit. The belief that HRT was beneficial was strongly held by both patients and physicians. Three trials examined the effect of HRT on cardiovascular disease [19–21]. In each of these trials, a negative harmful trend emerged for the primary outcome of cardiovascular mortality and morbidity. If these negative trends continued on to establish harm, these results would have an important impact on the continued use of HRT for reducing cardiovascular risk. On the other hand, if HRT is not beneficial, or at least non-inferior, then HRT would still be of clinical value and interest since it is effective for both improving symptoms and reducing bone loss. Given the strongly held beliefs about the benefits of HRT for cardiovascular disease, only convincing evidence of harm might change medical practice. Thus, each of these trials continued until there was sufficient evidence of harm, especially events related to the blood clotting effect of HRT. While the results of the first trial, HERS, had modest impact, all three trials together did reduce the amount of HRT usage for cardiovascular risk [22].

These examples illustrate the need to quantify the probability or likelihood that these interventions showing a negative trend during interim analyses would ultimately demonstrate a beneficial effect in the primary outcome. Some of the methods for those calculations of futility are described in the next section, using the VEST data for illustration.

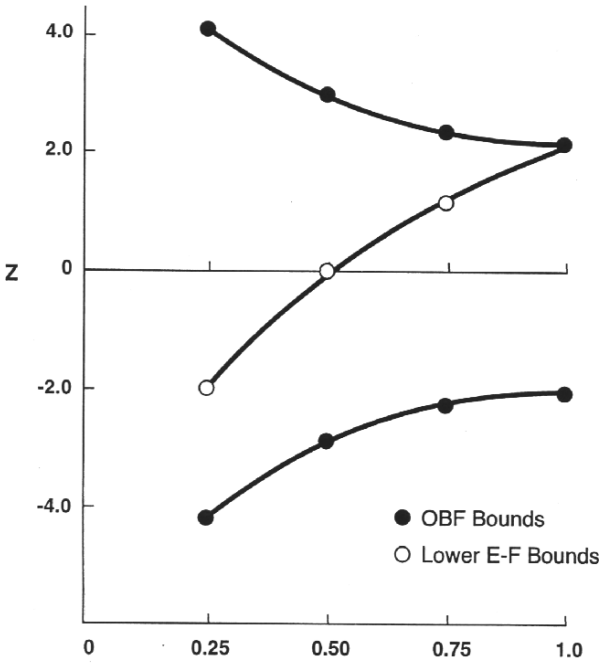
Futility analysis methods

One of the earliest trials to utilize the calculation of the likelihood of a positive benefit given a negative trend was the CDP which terminated several arms due to disappointing negative trends [23]. Since the CDP, statisticians have developed and refined methods for this type of futility analysis. We shall summarize briefly, four of these; beta spending functions, sequential probability ratio tests or triangular regions, conditional power, and a Bayesian method called predictive power, using the VEST data for illustration. As shall be noted, these methods are related in their theoretical basis and in operational characteristics.

Beta spending functions

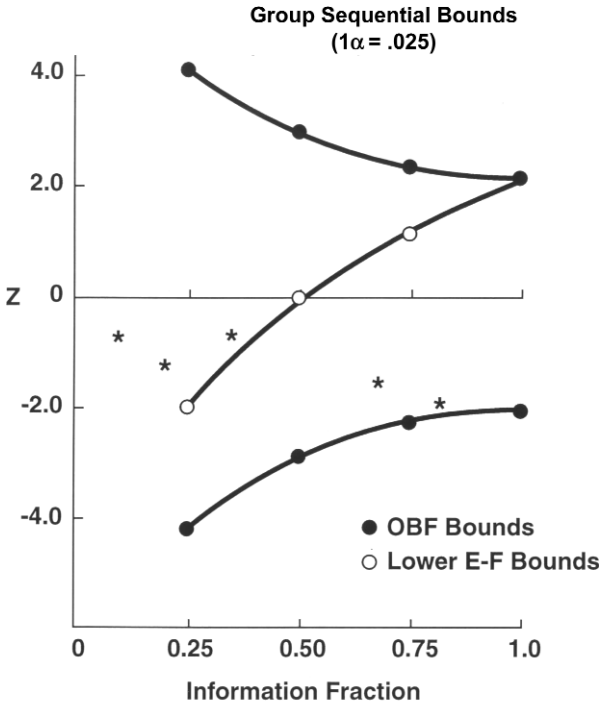
The group sequential methods for suggesting benefit or harm can be developed by using a method referred to as alpha spending functions [24]. These methods control the false positive or beneficial claim (Type I error) or alpha level for the repeated testing of accumulating data on the primary outcome, while allowing for flexibility in when and how many times the interim analyses are conducted. The beta spending function method is a similar approach but has the goal of ruling out a treatment effect of a prespecified size, which may be related to the alternative hypothesis used in the trial design [25,26]. Beta spending functions control the Type II or beta error. One such set of boundaries are shown in Figure 2, using an alpha spending function of the O'Brien-Fleming type for benefit (upper boundary, alpha = 0.025) or harm (lower boundary, alpha = 0.025) and a beta spending function (beta = 0.025) to rule out a benefit of a size of the hypothesized alternative effect.

The accumulating data (Table 1) are plotted in Figure 3 for the VEST trial [18], where the negative trends cross the beta spending function boundary midway through the trial. This beta spending function crossing indicates that the probability of a beneficial effect is less that 2.5%. However, with the



OBF = O'Brien–Fleming Bounds [12].
E-F = Emerson–Fleming Bounds [26].

Figure 2 Asymmetric beta spending function



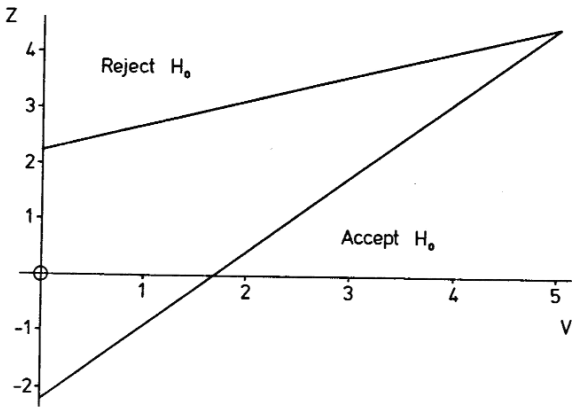
OBF = O'Brien–Fleming Bounds [12].
E-F = Emerson–Fleming Bounds [26].

Figure 3 Vesnarinone trial (VEST) group sequential bounds

established positive effects on some aspects of heart function and quality of life, the impressive mortality reduction in an earlier trial [17], and with an observed improved quality of life in VEST as well, the DMC allowed the trial to continue. Even if Vesnarinone demonstrated no beneficial effect on mortality, the benefits on quality of life would still remain of interest for heart failure patients. Thus, distinguishing between no beneficial effect versus harm on mortality was important.

Sequential probability ratio tests

Another methodology for creating boundaries to assess benefit or harm or lack of benefit is based on classical sequential probability ratio tests [27]. Figure 4 depicts such a boundary. The scale for this type of boundary is the numerator of the standard test statistic (eg, difference in proportions) on the vertical axis versuss. the variance of that difference on the horizontal axis. When plotted on the same scale, the group sequential boundaries using alpha and beta spending functions produce boundaries very similar to that of the sequential probability ratio tests (SPRT) [28]. Thus, while these boundaries at first may appear to be different, they are quite similar when using the same scale and thus have similar operating characteristics.



Reprinted from *Biometrics*, Whitehead & Stratton: Group Sequential Clinical Trials with Triangular Continuation Regions 1983; 39: 227–36, with permission from IBS.

Figure 4 Sequential probability ration test triangular boundaries

Conditional power

One of the earliest applications of the concept of conditional power was in the CDP [9,23]. In this trial, several treatment arms for evaluating cholesterol lowering drug produced negative trends in the interim results. Through simulation, the probability of achieving a positive or beneficial result was calculated given the observed data at the time of the interim analysis. The type of calculation is now often referred to as conditional power. Unconditional power is the probability at the beginning of the trial of achieving a statistically significant result at a prespecified alpha level and with a prespecified alternative treatment effect. Ideally, trials should be designed with a power of 0.80 to 0.90 or higher. However, once data begin to accumulate, the probability of attaining a significant result increases or decreases with emerging positive or negative trends. Calculating the probability of rejecting the null hypothesis of no effect once some data are available is conditional power, in contrast to power which is calculated at the beginning of the trial with no data yet available.

The calculation of conditional power is quite simple and straightforward [29–30]. If Z(t) represents the standardized statistic at information fraction t, where information fraction may be defined at the proportion of expected patients or events observed so far, then the conditional power for some alternative intervention effect θ is as follows, using a critical value of $Z\alpha$ for a Type I error of alpha:

$$P[Z(1) \geq Z_\alpha \mid Z(t), \theta]$$
$$= 1 - \Phi\{|Z_\alpha - Z(t)\sqrt{t} - \theta(1-t)|/\sqrt{(1-t)}\}$$

where $\theta = E\ (Z(t = 1))$ and where θ is defined as follows for:

1) Survival outcome (D = total events)

$$\theta = \sqrt{D/4}Ln(\lambda_C / \lambda_T)$$

where λ_C and λ_T are the hazard rates in the control arm and treatment arm respectively.

2) Binomial outcome (n = total sample size)

$$\theta = \frac{P_C - P_T}{\sqrt{2\bar{p}(1-\bar{p})/(n/2)}} = \frac{(P_C - P_T)\sqrt{N/4}}{\sqrt{\bar{p}(1-\bar{p})}} = 1/2\frac{(P_C - P_T)\sqrt{N}}{\sqrt{\bar{p}\bar{q}}}$$

where P_C and P_T are the event rates in the control arm and treatment arm respectively and $\bar{p}$ is the common event rate.

3) Continuous outcome (means) (n = total sample size)

$$\theta = \left(\frac{\mu_C - \mu_T}{\sigma}\right)\sqrt{N/4} = 1/2\left(\frac{\mu_C - \mu_T}{\sigma}\right)\sqrt{N}$$

where μ_C and μ_T are the mean response levels for the control arm and the treatment arm, respectively, and σ is the common standard deviation.

If we specify a particular value of the conditional power which would suggest futility, then a boundary can also be produced. As shown in Figure 5, the lower boundary is based on a specified conditional power that would be used to claim futility of a positive beneficial claim at the end of the trial. Plots are shown for 10%, 20% and 30% conditional power. For example, anytime the standardized statistic crosses that 20% lower boundary, the conditional power for a beneficial result at the end of the trial is less than 0.20 for the specified alternative intervention effect.

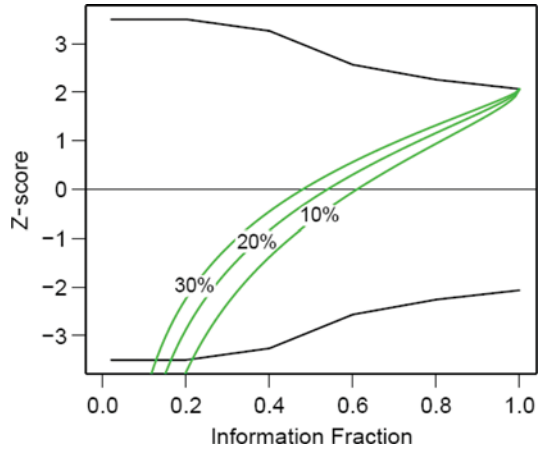


Figure 5 Conditional power boundaries: outer boundaries represent symmetric O'Brien–Fleming type sequential boundaries ($\alpha = 0.05$). Three lower boundaries represent boundaries for 10%, 20% and 30% conditional power to achieve a significant ($P < 0.05$) result of the trial conclusion

Typically, these conditional power calculations are done for a series of specified alternatives, ranging between the null hypothesis of no effect and the prespecified design based alternative treatment effect. In some cases, even more extreme beneficial effects can be utilized to determine just how much more effective the treatment would have to be to raise the conditional power to desired levels. If these results are summarized in a table or a graph, DMC members can assess for themselves whether they believe recovery from a substantial negative trend is very plausible. Table 2 provides conditional power for the VEST trial at three of the interim analyses for a range of intervention effects including the beneficial effect (hazard rate less than 1) seen in the previous trial to the observed negative trend (hazard rates of 1.3 and 1.5). As shown, the conditional power for a beneficial effect is very low (<0.01) by the midpoint of this trial (information fraction of 0.5) for a null effect or worse, that is, a higher relative risk, and not encouraging even for the original assumed effect ($RR = 0.7$).

Predictive power

One Bayesian method to assess futility is referred to as predictive power and related to the concept of conditional power [31,32]. In this case, the series of possible alternative treatment effects θ are represented by a prior distribution for θ, distributing the probability across the alternatives. The prior probability distribution can be modified by the current trend to give an updated posterior for θ. The conditional power is calculated as before for a specific value of θ. Then as shown below a predictive or 'average' power is calculated by integrating the conditional power over the posterior distribution for θ. This predictive power can then be utilized by the DMC to assess whether the trial is still viable:

Table 2 Conditional power for VEST*

RR†	Information fraction		
	0.50	0.67	0.84
0.50	0.46	<0.01	<0.01
0.70	0.03	<0.01	<0.01
1.0	<0.01	<0.01	<0.01
1.3	<0.01	<0.01	<0.01
1.5	<0.01	<0.01	<0.01

*VEST = Vesnarinone Trial [18].
†RR = relative risk.

Table 3 Predictive probability

		Probability	
Date	*t*	HR‡ = 0.40	Diffuse prior
2/7/96	0.50	0.28	0.0001
3/7/96	0.60	0.18	<0.0001
4/10/96	0.67	<0.0001	<0.0001
5/19/96	0.84	<0.0001	<0.0001
6/26/96	0.90	<0.0001	<0.0001

‡HR = hazard rate.

$$p(X_f \in R)|x_0) = \int p(X_f \in R|\theta)\ p(\theta|x_0)d\theta \cdot$$

We have computed this predictive power for the various interim analyses conducted in VEST and results are shown in Table 3. In this case, the prior was taken from the first vesnarinone trial where the observed reduction in mortality was over 60% ($RR = 0.40$). For these calculations, the prior was first set at the point estimate of the hazard ratio equal to 0.40 with a variance related to the size of the first trial. A second calculation was made with a diffuse prior. Of course, other distributions around the prior effect size could also have been used but the essence of this approach would be the same. Using this approach, it is clear that VEST will not show a beneficial effect since the probabilities in Table 3 are <0.001 for the latter stage of the trial. Thus continuation at those times might be termed futile.

Conclusions

All of the statistical methods summarized here have proven to be useful in monitoring trials with emerging negative trends. These methods generally do not differ in the message that they provide to the data monitoring committee and investigators. These methods can also be applied to ruling out a margin of non-inferiority in a similar fashion. However, the real challenge remains whether a trial with a negative trend should terminate early ruling out either a beneficial effect, or a margin of non-inferiority, or should continue in order to distinguish between this type of futility or harm. The answer to that question is often very difficult, and is not statistical, but depends on other beneficial aspects of the intervention. These aspects include whether the intervention is new or already established for other indications, and whether the intervention is already in practice for this indication. In addition, within the trial, other clinically relevant outcomes may be indicating positive beneficial trends. While the statistical methods are often very useful, the ultimate recommendation to terminate or continue depends largely on the judgment of a data monitoring committee and the initial guidance provided by the trial steering committee of investigators and sponsors.

References

1. **US Food and Drug Administration**. Draft guidance for clinical trial sponsors on the establishment and operation of Clinical Trial Data Monitoring Committees. Rockville, MD: FDA (http://www.fda.gov/cber/gdlns/clindatmon.htm), 2001.
2. **Committee for Medicinal Products for Human Use (CHMP) Guidelines**. Guidelines on Data Monitoring Committees, European Medicine Agency, July 2005.
3. **National Institutes of Health**. NIH policy for data and safety monitoring. NIH Guide (http://grants2.nih.gov/grants/guide/notice-files/not98-084.html), 1998.
4. **Ellenberg S, Fleming T, DeMets D.** *Data monitoring committees in clinical trials: A practical perspective*. John Wiley & Sons, Ltd., 2002.
5. **Greenberg Report**. Organization, review, and administration of cooperative studies. *Control Clin Trials* 1988; **9**: 137–48.
6. **Shalala D.** Protecting research subjects–what must be done. *New Engl J Med* 2000; **343**: 808–10.
7. **Psaty BM, Furberg CD.** COX-2 inhibitors–lessons in drug safety. *New Engl J Med* 2005; **325**: 1133–35.
8. **Armitage P, McPherson CK, and Rowe BC.** Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society Series A* 1969; **132**: 235–44.
9. **Canner PL, Klimt CR.** Experimental design features of the Coronary Drug Project. *Control Clin Trials* 1983; **4**: 313–32.
10. **DeMets DL, Lan KKG.** Interim analysis: The alpha spending function approach. *Stat in Med* 1994; **13**: 1341–52.
11. **Jennison C, Turnbull BW.** Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science* 1990; **5**: 299–317.
12. **O'Brien PC, Fleming TR.** A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**: 549–56.
13. **DeMets DL, Pocock S, Julian DG.** The agonizing negative trend in monitoring clinical trials. *Lancet* 1999; **354**: 1983–88.
14. **CONSENSUS Trial Study Group**. Effects of enalapril on mortality in severe congestive heart failure: Result of the Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS). *New Engl J Med* 1987; **316**: 1429–35.
15. **Budaj A, Herbaczynska-Cedro K, Kokot F *et al.*** Effect of early captopril treatment on blood adrenaline levels in acute myocardial infarction (the substudy of ISIS-4). International Study of Infarct Survival-4. *Amer J of Cardiol* 1998; **81**: 335–39.
16. **Packer M, Carver JR, Rodeheffer RJ *et al.*** Effect of oral milrinone on mortality in severe chronic heart failure. The PROMISE Study Research Group. *N Engl J Med* 1991; **325**: 1468–1475.
17. **Feldman AM, Bristow MR, Parmley WW *et al.*** for The Vesnarinone Study Group. Effects of vesnarinone on morbidity and mortality in patients with heart failure. *New Engl J Med* 1993; **329**: 149–55.
18. **Cohn J, Goldstein S, Greenberg B *et al.* for the Vesnarinone Trial Investigators**. A dose-dependent increase in mortality with vesnarinone among patients with severe heart failure. *New Engl J Med* 1998; **339**: 1810–16.
19. **Grady D, Applegate W, Bush T *et al.*** Heart and estrogen/progestin replacement study (HERS): design,

methods, and baseline characteri-stics. *Control Clin Trials* 1998; **19**: 314–35.
20. **Writing Group for the Women's Health Initiative Investigators**. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. Principal results from the Women's Health Initiative Randomized Clinical Trial. *JAMA* 2002; **288**: 321–33.
21. **The Women's Health Initiative Steering Committee**. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy. The Women's Health Initiative Randomized Controlled Trial. *JAMA* 2004; **291**: 1701–12.
22. **Hulley SB, Grady D.** The WHI estrogen-alone trial - things look any better? Editorial. *JAMA* 2004; **291**: 1769–71.
23. **Coronary Drug Project Research Group**. Clofibrate and niacin in coronary heart disease. *JAMA* 1975; **231**: 360–81.
24. **Lan KKG, DeMets DL.** Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**: 659–63.
25. **Emerson SS, Fleming TR.** Symmetric group sequential test designs. *Biometrics* 1989; **45**: 905–32.
26. **Tsiatis AA, Rosner GL, Mehta CR.** Exact confidence intervals following a group sequential test. *Biometrics* 1984; **40**: 797–803.
27. **Whitehead J, Stratton I.** Group sequential clinical trials with triangular continuation regions. *Biometrics* 1983; **39**: 227–36.
28. **Whitehead J.** A unified theory for sequential clinical trials. *Stat Med* 1999; **18**: 2271–86.
29. **Lan KKG, Wittes J.** The B-value: A tool for monitoring data. *Biometrics* 1988; **44**: 579–85.
30. **Lan KKG, Zucker D.** Sequential monitoring of clinical trials: The role of information in Brownian motion. *Stat Med* 1993; **12**: 753–65.
31. **Spiegelhalter DJ, Freedman LS, Blackburn PR.** Monitoring clinical trials: conditional or predictive power? *Control Clin Trials* 1986; **7**: 8–17.
32. **Spiegelhalter DJ, Freedman LS, Parmar MKB.** Bayesian approaches to randomized trials. *J Royal Statist Soc A* 1994; **157**: 357–416.