Monitoring of Clinical Trials:
Issues and Recommendations

Thomas R. Fleming, PhD and David L. DeMets, PhD

Department of Biostatistics, University of Washington, Seattle (T.R.F.) and Department of Biostatistics, University of Wisconsin, Madison (D.L.D.)

ABSTRACT: Interim analyses of randomized trials enable investigators to make more efficient use of limited research resources and to satisfy ethical requirements that a regimen be discontinued as soon as it has been established to have an inferior efficacy/toxicity profile. Unfortunately, the integrity and credibility of these trials can be compromised if inappropriate procedures are used in monitoring interim data. In this paper we discuss how group sequential designs provide useful guidelines that enable one to satisfy the valid objectives of interim monitoring while avoiding undesirable consequences, and we consider how flexible one can be in the way such designs are implemented. We also provide motivation for the role of data-monitoring committees in preserving study integrity and credibility in either government- or industry-sponsored trials. In our view, these committees should have multidisciplinary representation and membership limited to individuals free of apparent significant conflict of interest, and ideally should be the only individuals to whom the data analysis center provides interim results on relative efficacy of treatment regimens. Finally, we discuss some important practical issues such as estimation following group sequential testing, analysis of secondary outcomes after using a group sequential design applied to a primary outcome, early stopping of negative trials, and the role of administrative analyses.

KEY WORDS: *group sequential designs, interim analyses, data-monitoring committees, administrative analyses, active control trials, repeated confidence intervals, stochastic curtailment*

INTRODUCTION

Frequently results from clinical trails are monitored over time. This practice raises a number of important issues concerning proper conduct and interpretation of interim analyses in these prospective clinical studies. We will discuss many of these issues, review some relevant research, and provide recommendations. We begin with a review of the history of and motivation for trial monitoring.

Interim analyses of clinical trial data are strongly advised, indeed ethically mandated, to assess whether early evidence either convincingly establishes treatment benefit or convincingly establishes that treatment will not provide clinically meaningful benefit. These basic principles were set down in 1967

Address reprint requests to: Thomas R. Fleming, PhD, Department of Biostatistics, University of Washington, Seattle, Washington 98119.

Received October 7, 1991; revised December 1, 1992.

Controlled Clinical Trials 14:000–000 (1993)

655 Avenue of the Americas, New York, New York 10010

0197-2456/93/$6.00

by an NIH National Heart Institute committee chaired by Dr. Bernard Greenberg. The committee's report, known as the Greenberg Report, was recently published in *Controlled Clinical Trials* [1]. The principles of this report have been adopted by many NIH-funded multicenter clinical trials [2,3], especially in the areas of cardiovascular disease, cancer, ophthalmology, and AIDS, since the report was originally issued.

These interim analyses allow data from the trial itself to be used in decisions about whether, for ethical, scientific, or economic reasons, the study should be terminated early, i.e., prior to the protocol-specified time of trial completion. Patient, health care personnel, and fiscal resources are precious and must be used properly. It is well recognized that interim analyses of randomized clinical trials enable investigators to make more efficient use of limited resources in patients, health care systems, time, and dollars [2]. Ethically, trials can only be initiated and continued when it is not known which of two treatments or strategies is superior or inferior. A basic principle is that clinical trials will not be conducted longer than necessary to reach the goal of the trial. Early evidence may strongly suggest that either benefit or lack of meaningful benefit has already been established and it would be unethical to add more patients or continue follow-up without stopping the trial and sharing that information with current and prospective patients.

Trials may also be terminated early for practical reasons. For example, the design assumptions may not be consistent with observed baseline data. The design may have depended on a higher risk patient than is actually being recruited, thereby rendering the trial underpowered, unless modified. In addition, economic issues may play a role in early termination. Perhaps the patient resources have been overestimated, projected costs to conduct a trial may not have been realistic, or patient recruitment may have been slower than anticipated. To achieve the original goal would require a considerable extension of effort, time, or dollars, perhaps beyond the available fiscal resources. Issues such as these do not involve primary outcome results and are considered administrative.

When interim monitoring of trials is performed, one does need to be cautious when interpreting the strength of evidence about treatment effects. Monitoring of data without proper adjustment for taking repeated looks will lead to a substantial increase in the likelihood of obtaining false-positive or false-negative conclusions. This is a well-known issue and has been described, for example, in work by Armitage and colleagues [4]. If a study uses a critical value of 1.96 (i.e., two-sided $P \leq .05$), then for a single analysis the false-positive error is .05 as designed. However, suppose this 1.96 critical value is used at interim analyses performed after equal increments of information. For two analyses, the false-positive error rate would be .08, for five analyses .14, and for 10 analyses almost .20. It has also been shown [5,6] that if survival data are monitored four times a year over a 4-year period for a typical cancer clinical trial, then one quarter of the trials will yield a logrank P value $<.05$ at some point in time. The Coronary Drug Project [7] illustrated what *might* have happened if investigators had monitored the clofibrate–placebo mortality comparison using 1.96 as the critical value. As shown in that paper, this value was reached on several instances, yet the ultimate mortality curves were

sevier—Controlled Clinical Trials—04-07-93 14-38-57
CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613

nearly superimposed. Subset analyses conducted at interim analyses [8] can further increase the likelihood of reaching false conclusions.

It is apparent that statistical methods to guide the interpretation of interim results would be useful. In the next section we discuss how group sequential designs provide guidelines that enable one to satisfy the objectives of interim monitoring while avoiding some of the undesirable consequences, and we consider how flexible one can be in the way such designs are implemented. Some of the important practical issues arising during interim monitoring are considered in a later section when we discuss our perspective on the structure and purpose of data-monitoring committees. We review a number of issues that should be considered when defining the composition and function of such committees. Following that, we discuss some selected topics, such as estimation following group sequential testing, analysis of secondary outcomes after using a group sequential design applied to a primary outcome, analysis of active control designs, early stopping of negative trials, and the role of administrative analyses.

GROUP SEQUENTIAL DESIGNS: THEIR ROLE AND FLEXIBILITY

Role of Group Sequential Guidelines

Interim data analyses for large-scale multicenter clinical trials have been in place for more than 20 years. The Coronary Drug Project was among the first to take into account the issue of repeated testing and to describe the decision-making process [7]. Because this process is complex, no simple stopping rules can be stated. However, useful guidelines have been developed in recent years. DeMets [9], Fleming and Watelet [8], Jennison and Turnbull [10], and Emerson and Fleming [11] have presented reviews of statistical methods for interim analyses and discuss approaches to designing trials that allow investigators to adjust their estimation and testing procedures when multiple looks at the data are planned. Two basic methods are the group sequential approach and the stochastic curtailment approach.

The group sequential approach, proposed by Pocock [12] and O'Brien and Fleming [13], is one in which plans are made for a small number of interim analyses (e.g., two to eight), in contrast to classical sequential methods that call for analyses after the recording of each outcome. This alternative to classical sequential methods is motivated in part because performing more than a few interim analyses provides little additional increase in efficiency and because data management constraints usually do not allow for the continued availability of high-quality data. Both the Pocock and the O'Brien–Fleming procedures require that interim tests be performed at conservative levels (i.e., a larger critical value for statistics to be judged significant or, equivalently, a smaller nominal P value) to avoid obtaining excess false-positive or false-negative conclusions. The merits of group sequential procedures have been described by DeMets [14]. Others, such as Peto et al. [15] and Haybittle [16], have proposed related monitoring approaches that also suggest the use of conservative levels for interim tests.

The O'Brien–Fleming group sequential design has several important prop-

Elsevier—Controlled Clinical Trials—04-07-93 14-38-57
93CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613

erties, as discussed by Simon [17]. It is very conservative at early analyses when results are likely to still be unreliable, it does not require an increase in the protocol-specified total sample size to approximately maintain the power of a fixed sample trial, when trials go to scheduled completion it requires very little adjustment to significance levels obtained at final analyses. One of the first implementations of this design was in the Beta-Blocker Heart Attack Trial [18,19], which in fact was stopped early because of a beneficial effect. Many other trials in which it has been used to stop early include the NCI Cancer Intergroup study no. 0035 investigating 5-FU + levamisole in the colon adjuvant setting [20], the NIAD AIDS Cooperative Study Group trials nos. 016 and 019 providing controlled evaluation of AZT in early ARC, and in asymptomatic patients, respectively [21], and an industry-sponsored trial of γ-interferon in chronic granulomatous disease [22]. A group sequential procedure with similar attributes was used in the recent Cardiac Arrhythmia Suppression Trial (CAST) [23].

DeMets [9] reviewed experiences of interim monitoring in noncancer trials, while Rosner and Tsiatis [24] investigated the impact of group sequential designs on cancer cooperative group clinical trials and have found that use of these designs "could produce a substantial saving in time and, hopefully, patient failures." Green and Fleming [25] discussed group sequential guidelines for interpretation of repeated looks at maturing data that accrue after the time of early termination (e.g., the current situation for Cancer Intergroup Study no. 0035 investigating 5-FU + levamisole in the colon adjuvant setting). These authors found that periodically applying conservative significance criteria to such maturing data would be an effective approach to maintaining low false-positive and false-negative error rates. As will be discussed later, DeMets and Ware [26] and Emerson and Fleming [27] considered extensions of Pocock or O'Brien–Fleming designs in order to provide guidelines for early stopping of negative as well as positive studies.

The second major approach used in data monitoring is stochastic curtailment. Initially, this concept was referred to as conditional power since the basic idea is to compute the probability of rejecting the null hypothesis, given that part of the trial is already completed, as illustrated by Halperin et al. [28]. A more formal theoretical structure was provided by Lan, et al. [29], with a method for calculation by Lan and Wittes [30]. Suppose that this conditional probability of rejecting the null hypothesis at the scheduled or planned termination point, given the interim results already available, is high even if we assume there is no true treatment effect for the remainder of the trial. We might then consider terminating the trial since with high probability the final conclusion is already "known." For early termination, if this probability were required to be over .90, then the increase in the type I error would be small [29]. This idea was used, for example, in the Beta-Blocker Heart Attack Trial [18]. For either a negative (i.e., harmful) trend or a lack of trend, we might compute the probability of rejecting the null hypothesis of no treatment effect under a variety of reasonable alternatives, including the treatment effect proposed in the study protocol. If this probability is small, we might then "know" that there is little chance of claiming a treatment benefit at trial completion. This concept was used in part in terminating the CAST given

Elsevier—Controlled Clinical Trials—04-07-93 14-38-57
93 CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613

the early strong negative trend. A Bayesian version of this concept has been described by Spiegelhalter et al. [31].

Flexibility in Implementing Group Sequential Designs

Group sequential designs should not be used to provide stopping rules per se. Rather they are guidelines as discussed by the Coronary Drug Project Research Group [7]. Fleming et al. [32], DeMets [9], and Meier [33]. Ideally, recommendations regarding early termination of trials should be made by data-monitoring committees (DMCs) having multidisciplinary representation. In our view, their recommendations should be guided by group sequential designs yet should be global in that all available information must be taken into account. For example, consider the recently reported colon adjuvant trial of 5-FU + levamisole [20]. The DMC for that trial recommended that the study be terminated at the second stage of a four-stage O'Brien–Fleming design when the logrank statistic for the association of treatment with patient survival yielded a two-sided $P = .006$, lower than the O'Brien–Fleming guide, $P = .01$. However, the DMC might have judged the trial to be conclusive, even if a somewhat higher P value for survival had been observed, due to striking delays in time to recurrence and due to supportive data from an earlier trial with an identical design. Interestingly, the DMC had received considerable pressure in the press [34] to stop the trial much earlier.

Interim analyses do not need to be performed after equal increments of information are obtained (e.g., after equal increments of deaths in a survival analysis), and the original O'Brien–Fleming [13] paper provides formulas for adjustment when one performs testing after unequal increments. DeMets and Gail [35] also showed that for the logrank test the types I and II errors are not appreciably affected if equal-increment group sequential plans are implemented even though analyses are done at unequal increments.

Given that the decision by a DMC to recommend trial termination involves a complex process and is not simply reached by application of "cut-and-dried" stopping rules, the DMC may decide in the best interests of the patients to review the data more or less frequently than specified by the original group sequential design. The CAST [23] is a case in point. An important issue relates to how flexible we can be, once a trial is underway, in changing the timing of analyses and the maximum number of looks to be taken. Lan and DeMets [35] showed that the effect of such changes on false-positive and false-negative error rates is small as long as a use function has been specified in advance for the trial. The use function, conceptually proposed by Lan and DeMets [37], involves specifying the rate at which the false-positive error will be spent as a function of the proportion of total information achieved. For example, specifying that an O'Brien–Fleming or a Pocock design is to be employed does put a well-defined use function in place. If one plans a trial with four interim analyses using an O'Brien–Fleming guideline, during the study one could double the number of analyses in the O'Brien–Fleming design without meaningfully inflating the error rates, as indicated by Lan and DeMets [36]. On the other hand, changing the use function in midstream could have serious consequences on error rates. As a trivial illustration, if one monitors data

Elsevier—Controlled Clinical Trials—04-08-93 08-09-04
93CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613

continuously and decides to spend all the false-positive error the first time a $P < .05$ is achieved, contrary to what a proper use function would allow, the same excessive false-positive error will arise that was illustrated by Fleming [5].

It is important to have some guidelines in place in advance in clinical trials. This should be considered by those involved in protocol development, including those in the Food and Drug Administration (FDA) or other regulatory agencies who are guiding or approving study designs from industry. In our experience, group sequential designs implemented via use functions offer an attractive and helpful approach, providing the flexibility needed by a DMC.

Certain design assumptions may need reexamination in the early stages of the clinical trial to verify that available resources will be adequate to meet study objectives. We may discover that the risk level or accrual rate of the subjects is less than anticipated, so that the power to detect the hypothesized effect is compromised. Thus, the sample size may need to be increased to recover desired levels of power. If we fix the required number of events in the design of a survival study, we may use early trial information to determine the increase in the number of subjects needed to obtain that number of events. This determination should be made before unblinding results on the relative efficacy of treatments. Group sequential designs will not protect the type I error if relative efficacy results are used in making decisions about sample size changes.

THE DATA-MONITORING COMMITTEE: STRUCTURE AND PURPOSE

As noted earlier, interim analyses of clinical trials allow one to monitor for early extreme therapeutic results as well as for excess toxicity and practical difficulties. However, routine broad reporting of interim results could be detrimental to study integrity. Furthermore, recommendations about trial design alterations or early termination, if made by individuals having apparent conflicts of interest, could be detrimental to study credibility.

In our view, to effectively preserve study integrity and credibility and to safeguard the rights of patients, independent DMCs should be established for all pivotal randomized trials, i.e., those trials designed to enable definitive assessments of the therapeutic effects of interventions. This is particularly important in the setting of diseases that are life threatening or provide irreversible morbidity. Several issues should be considered in defining the composition and function of these committees. The DMC:

1. Should have multidisciplinary representation including physicians from relevant medical disciplines and biostatisticians, and often should have other experts such as ethicists or epidemiologists.
2. Should have membership limited to individuals free of apparent significant conflicts of interest, whether they be financial, scientific, or regulatory in nature.
3. Should be ethically and scientifically supportive of study objectives and design.
4. Should balance its responsibilities to three groups of patients: those already enrolled onto the study, those yet to be enrolled, and future patients outside the study.

Elsevier—Controlled Clinical Trials—04-08-93 08-09-04
93CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613

5. Should be knowledgeable about available external information that is relevant to its decision-making responsibilities.
6. Should be aware of data management and quality control procedures employed in the trial, and should be confident that the committee has access to accurate and complete data.
7. Should be guided by protocol-specified group sequential designs as well as by an unblinded broad overview of all relevant available results when making decisions about whether early termination should occur.
8. Should ideally be the only individuals to whom the data analysis center provides interim results on relative efficacy of treatment regimens.
9. Should consider having open sessions during which the committee can be provided information by industry/government sponsors, study investigators/statisticians, or the FDA, and closed executive sessions at which data on the relative efficacy of treatments are discussed.
10. Should have procedures to evaluate and act on special requests from study investigators or sponsors to provide them limited access to some evolving study information. These procedures should not unblind non-DMC members to relative efficacy results.
11. Should independently make its recommendation to continue or terminate a trial to the sponsoring agency (e.g., NIH, industry) or to study investigators, taking into account safety and efficacy results.

It might be helpful to provide more specific motivation for this formulation of the composition and function of the DMC. Due to the complexity of clinical trials and the decision-making process, the committee should have sufficiently broad multidisciplinary representation to ensure that all relevant ethical, safety, medical, and scientific issues can be adequately discussed and properly weighed in all recommendations concerning trial conduct and termination.

Study integrity and credibility are compromised if decisions about whether to terminate a trial early are influenced by individuals having apparent conflicts of interest. This could occur, for example, if the study sponsor prematurely withdraws support in order to achieve objectives in conflict with the need to scientifically address trial objectives, or if the sponsor or study investigators should attempt to manipulate the conduct of the trial or interpretation of its results to achieve financial or scientific benefit. Unfortunately, our experience has shown that these problems can occur in trials lacking proper DMCs. Due to these concerns, sponsors or other individuals having significant financial or professional interests dependent on the outcome of the clinical trial should not be members of the committee. DMC members should disclose relevant financial interests as well as other types of apparent significant conflicts, as recently endorsed by Healy et al. [38].

As noted by the Coronary Drug Project [7] and Green et al. [39], judgments about whether to continue a clinical trial should weigh responsibilities to three groups of patients: those already enrolled in the study, those yet to be entered and future patients outside the study. Premature termination of a trial can produce significant negative consequences for each of these groups. Quoting the latter authors:

> The commitment and cooperation of patients currently on study are wasted if

a study becomes equivocal or misleading. Thousands of future patients are at risk for receiving an ineffective or costly or toxic treatment (if a treatment is erroneously reported as superior) or are at risk for not receiving an effective treatment (if a new treatment is erroneously reported as being not better than a standard). Even patients yet to be entered on study, the very patients we most seek to protect by early termination, are not necessarily helped by such action, since they are likely to receive the regimen that appeared preferable in early data, even though more data or larger follow-up might have shown it to be inferior.

In our view, the responsibilities to all patients are best served by charging well-informed DMCs with the responsibility for making recommendations for early termination.

Interim monitoring of trial results does provide considerable demands on the data management resources. Since any interim analysis could lead to termination of the trial, complete and accurate data must be available to the DMC at each time of analysis. The DMC should review data management and quality control procedures employed in the study and be confident that the key efficacy and safety data are complete and accurate whenever an interim analysis is performed. In order to maximize available information and reduce the risk that subsequent data updates would substantively alter analysis conclusions, nearly current follow-up should be available on all patients. Specifically, in our view, a lag of more than 2 months between average patient last contact date and meeting date usually is not acceptable. Decisions about early termination should be delayed if available data do not provide nearly current follow-up on almost all patients.

Since it is common for early results to be misleading by giving the inaccurate impression that treatment effects are markedly favorable or unfavorable, broad reporting of interim therapeutic results greatly increases the risk of misinterpretation of what is reliably known about treatment effects. This increases the risk of inappropriate early abandonment of the trial. Green et al. [39] considered results from cancer cooperative groups and documented that when DMCs were employed there was a striking reduction in the number of trials showing a declining accrual rate over time, trials that were stopped early without meeting protocol specified objectives, and trials having early published results that were inconsistent with final results.

In pivotal trials designed to provide a definitive assessment of treatment effects, results on relative efficacy of treatments ideally should be available only to membership of the DMC and to individuals as the data monitoring and analysis center responsible for providing results to the committee. If efficacy results are known to individuals outside the committee, those individuals should be identified in advance and should agree to maintain the confidentiality of this information. However, in order to allow the committee to have adequate access to information provided by industry or government sponsors, by study biostatisticians or investigators, or by members of the FDA, a joint session between these individuals and committee members (called an open session) could be held to ensure that these important interactions do occur. Sessions involving only DMC membership (called closed sessions) could be held before and after the open session to allow discussion of data

Elsevier—Controlled Clinical Trials—04-08-93 08-09-04
93CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613

on the relative efficacy of treatments. Finally, the DMC should be responsive to the needs of sponsors, investigators, or regulators who, during the conduct of the study, are planning future studies, considering future steps in resource allocation or product development, or preparing for regulatory review. The DMC should have procedures to evaluate or act on special requests from such individuals to provide limited access to some evolving study information that does not require unblinding relative efficacy results.

In industry-sponsored trials, to assure proper masking, it is preferable to have the data management and analysis center as well as the DMC be independent of industry involvement. In the setting where logistical or financial considerations require that the data management and analysis center be "in-house," it is desirable that the only people from industry having access to relative efficacy results be individuals from the in-house data analysis center who are responsible for providing these results to the DMC. One compromise would be to have the data management center be in-house while providing for an independent data analysis center. This compromise was recently used in an industry cardiovascular trial [40].

SOME PRACTICAL ISSUES

In this section, we provide our views on several controversial issues that frequently arise in prospectively monitored clinical trials.

1. *Obtaining unbiased estimates of treatment effect and constructing confidence limits, once a group sequential trial has been completed.* Just as repeated analysis of accumulating data causes an increased risk for obtaining false-positive or negative conclusions, it also causes biased estimates of treatment effect. Whenever extreme results are observed, the trial is stopped. If observed data are not extreme, the trial is continued. Based on this, it is reasonable to expect the usual estimates appropriate for fixed sample designs to be biased toward the extremes that caused the study to be terminated early. This intuitive reasoning has been validated by rigorous studies that have shown that usual estimates of treatment effect have a 10–15% bias, while usual 90% confidence intervals can have coverage probabilities as low as 80% [41]. Such bias is worse when using Pocock-type designs than when using O'Brien–Fleming-type designs.

Methods have been proposed by Jennison and Turnbull [42], Tsiatis et al. [43], Whitehead [44], Chang and O'brien [45], and Kim and DeMets [46], Chang [47], and Emerson and Fleming [41] to provide essentially unbiased estimates of treatment effect as well as well as confidence intervals having correct coverage probabilities. Unlike the group sequential designs that have become routinely used as hypothesis-testing methods to guide early stopping decisions, these methods for obtaining unbiased estimates of treatment effect are just now beginning to be employed [11]. This is largely due to their greater complexity and more recent development, and the fact that necessary software is still being developed.

It is important to observe that the "repeated confidence intervals" proposed by Jennison and Turnbull [48] and illustrated in Fleming and Watelet [8] do not serve the same role as the confidence intervals, generated following com-

pletion of a group sequential trial, which were discussed in the previous paragraph. For illustration, at the second stage of an O'Brien–Fleming design, the interval in Fleming and Watelet is a fixed sample 99.6% confidence interval whose lower and upper limits drive decisions for early stopping of positive and negative studies, respectively. If the trial were to be stopped at this second analysis, one should not compute a 99.6% fixed sample confidence interval to represent the magnitude of treatment effect, but instead should use methods from the previous paragraph to obtain proper 95% confidence intervals.

2. *Obtaining adjusted significance levels and confidence intervals to assess the association of treatment with key secondary outcomes, in a completed trial that used a group sequential design applied to a primary outcome.* It is not uncommon for study investigators to base early stopping decisions for a clinical trial on a prespecified primary outcome, while there might be interest in evaluating treatment effect on key secondary outcomes once completed trial results have been reported. For example, as reported at the FDA Oncology Advisory Committee meetings held in February 1990 and in September 1990, early stopping occurred at the third of four planned analyses in a prospective clinical trial comparing experimental treatment with idarubicin + ARA-C vs. standard induction therapy with daunorubicin + ARA-C in ANLL. The O'Brien–Fleming boundary had been crossed by statistics evaluating treatment effect on the protocol-specified primary outcome measure, the rate of complete response. The FDA was interested in using these final results to evaluate the effect of treatment on a second efficacy measure, patient survival, and asked what adjustment would be required to account for the group sequential data evaluation.

Before sketching a rigorous approach to address this issue, it should be noted that the essence of the answer is intuitively clear. Returning to the illustration, if treatment's effect on the complete response rate is statistically independent of its effect on length of survival, then unadjusted significance levels and confidence intervals could be used in the analysis of treatment effect on patient survival. However, if treatment effect on complete response rates is highly correlated with or predictive of effect on survival, then nearly full group sequential adjustment would be required when evaluating the survival data. Thus the degree of adjustment is driven by this degree of correlation.

In current research to provide a rigorous solution to this problem, Emerson and Banks [49] formulate the joint distribution of the statistics for the primary and secondary outcomes, and then numerically integrate to find the distribution of the secondary outcome statistic after incorporating the stopping rule. This approach yields properly adjusted significance levels and confidence intervals for the key secondary outcome in much the same manner used for primary outcomes by Emerson and Fleming [41]. Earlier attempts at this problem were presented by Whitehead [44]. If several secondary outcomes are being considered, there are practical limitations for this approach.

3. *Early termination guidelines in trials having active control designs.* In trials with active control designs, the experimental treatment is compared to active standard therapy. The usual intent is to establish that the experimental treatment provides efficacy equivalent to that of the standard. This is in contrast to trials having a "no-treatment" control whereby one must show that the

experimental is superior to no treatment. However, by using confidence intervals to present relative efficacy results, the evaluation of treatment effect in an active control design can be performed in a manner conceptually parallel to the evaluation performed in a classical no-treatment control design. A detailed presentation of this confidence interval–based approach appears in Fleming [50]. These methods were used at FDA Oncology Advisory Committee meetings that considered mitoxantrone for advanced breast cancer in 1986, as described by Fleming [51], carboplatin for ovarian cancer in December 1988, and idarubicin for ANLL in February 1990.

Once treatment evaluation has been cast in terms of confidence intervals, group sequential guidelines can be applied to an active control study in the same manner done for the classical no-treatment control designs. Essentially repeated confidence intervals are constructed to identify what treatment differences can be ruled out with high probability. Once all differences of interest can be ruled out, early termination should be considered. Illustrations of the repeated confidence interval approach are presented in Fleming [50] and in Fleming and Watelet [8], with rigorous details presented by Emerson and Fleming [27] and Jennison and Turnbull [10,48].

4. *Group sequential designs or stochastic curtailment to guide early decisions about whether a trial is negative.* Clinical trails may be terminated early due to convincing evidence of lack of treatment benefit. The CAST trial serves to illustrate how group sequential designs and stochastic curtailment methods can be used. In CAST, two antiarrhythmia drugs were compared to a placebo. Early in the trial, the total observed mortality was 56:22 in favor of the placebo. Stochastic curtailment calculations indicated that the treatments would have to be much more effective than originally proposed, in contrast to the observed harmful effect, to reverse the negative trend and arise the probability of claiming benefit to desirable levels (e.g., over 80%). Since such a treatment benefit did not seem reasonable, even a priori, the probability was negligible for rejecting the null hypothesis and claiming the treatment superior. In addition, upper and lower group sequential boundaries were used that were symmetrical with respect to the hypothesis of no treatment effect; since no one expected a lower boundary for harm to be necessary, it was termed advisory. As it turned out, the lower "advisory" boundary was, in fact, crossed. This provided convincing evidence that not only was the treatment not beneficial, it was harmful.

In developing a lower boundary to guide early decisions about whether a trial is negative in a group sequential setting, authors have varied in whether they chose it to be asymmetrical or symmetrical relative to the upper boundary. DeMets and Ware [26] discussed the idea of asymmetrical boundaries where less stringent evidence might be used to declare harm or lack of treatment benefit than to claim treatment benefits. They proposed this asymmetry for a pediatric study. Canner [52] also discussed asymmetrical boundaries in the context of monitoring a trial with correlated primary outcomes.

Recently, Emerson and Fleming [27] explored symmetrical group sequential designs, which extend the concept of Pocock or O'Brien–Fleming designs, in order to provide guidelines that require the same strength of evidence for stopping negative trials as for positive trials. By these symmetrical designs, a trial is positive when early favorable results are highly inconsistent with

the hypothesis of no treatment effect, and is negative when early unfavorable results are highly inconsistent with the hypothesis of a clinically meaningful treatment benefit, where "highly inconsistent" is defined by the same type of group sequential boundary in both settings. In a trial like the CAST, for example, the symmetrical group sequential boundary would indicate the trial to be negative when early survival results are inconsistent with the smallest survival improvement judged to be clinically relevant, such as a 25% reduction in death rate for patients receiving an antiarrhythmia drug. This yields a criterion for negativity, easily satisfied by the actual CAST data, which is much less stringent than requiring definitive evidence that the drugs have a harmful effect on survival.

A related approach to interim analysis and decision making for negative as well as positive trials is the use of repeated confidence intervals as proposed by Jennison and Turnbull [48]. Confidence intervals are constructed at each interim analysis by using the group sequential boundary value as the confidence interval coefficient. As soon a the repeated confidence interval rules out all differences of interest, then the trial has the potential for early termination.

5. *Distinguishing characteristics of administrative analysis (not requiring statistical adjustment in the group sequential analyses) relative to formal interim analyses.* As already discussed, some analyses involve comparing treatment efficacy results, while other analyses do not use treatment outcome comparisons. We have designated any analyses comparing treatment outcomes as "interim analyses" and, in our view, adjustment for these analyses should be performed using a predefined group sequential or stochastic curtailment procedure.

"Administrative analyses" occur when one wishes to evaluate factors that could affect the integrity of the trial but that can be assessed without revealing relative efficacy results. Examples of such factors include comparability of patient characteristics across treatment regimens, baseline comparisons across clinics to detect recruitment differences, evaluation of baseline characteristics to assess risk levels and design assumptions, satisfaction of eligibility criteria, compliance to treatment regimens, and quality of data collection including completeness and accuracy of data. Decisions could be made to terminate a trial if the design is no longer viable, recruitment is hopelessly behind schedule, or data quality is unacceptable. To be categorized as administrative, these analyses must be conducted without access to data on the relative efficacy of treatments. This type of analysis does not affect the type I or II errors regarding the primary hypotheses.

SUMMARY

Pivotal phase III and phase IV trials typically require administrative and interim analyses to evaluate the progress of the trial in terms of ethical and scientific considerations. Through administrative analyses, investigators monitor the logistical and design aspects with the intention of achieving the best possible study. These analyses do not require comparisons of treatment group outcomes, be they primary or secondary variables. As such, no adjustments to usual significance levels are required. When outcome variables are being

Elsevier—Controlled Clinical Trials—04-07-93 14-38-57
93CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613

compared, we have referred to these as interim analyses. A major objective in interim analyses is to evaluate treatments and terminate trials early if treatment has already been established as beneficial, shows evidence of being harmful, or perhaps has no trend at all and differences of interest can be ruled out. Repeated analyses of an outcome variable with no adjustment can substantially increase the type I or II errors beyond acceptable levels. We have recommended two basic statistical methods which allow investigators to make proper adjustments when performing interim analyses, group sequential designs, and stochastic curtailment. While these monitoring guidelines are helpful, they do not incorporate all of the complexities of decision making in clinical trials. Due to this complexity, many multicenter trials sponsored by NIH and an increasing proportion of those sponsored by industry have used an independent DMC to integrate all relevant available information, including the results of group sequential or stochastic curtailment approaches, in evaluating whether or not a trial should continue.

We strongly encourage that future NIH and industry-sponsored pivotal trials, in particular those evaluating interventions in the setting of diseases that are life threatening or provide irreversible morbidity, use the independent DMC model as well as the statistical methods we have described. In turn, to enable readers to properly interpret data, authors reporting these study results should identify the type of statistical guidelines and DMC used to monitor the trial. This overall combination of statistical methods and decision-making committees has contributed substantially to preserving the integrity and credibility of many pivotal clinical trials.

REFERENCES

1. Greenberg Report: Organization, review, and administration of cooperative studies. Controlled Clin Trials 9:137–148, 1988
2. Friedman LM, Furberg CD, DeMets DL: Fundamentals of Clinical Trials. 2nd ed. Littleton, MA, PSG, 1985
3. Meinert CL, Tonascia S: Clinical Trials: Design, Conduct and Analysis. New York. Oxford University Press, 1986
4. Armitage P, McPherson CK, Rowe BC: Repeated significance tests on accumulating data. (Series A) 132:235–144, 1969
5. Fleming TR: Design consideration for clinical trials. In: Cancer Chemotherapy: Challenges for the Future, Vol. 3. K. Kimura et al. eds. Tokyo, Excerpta Medica, 1988
6. Fleming TR, Green SJ, Harrington DP: Considerations for monitoring and evaluating treatment effects in clinical trials. Controlled Clin Trials, 5:55–6, 1984
7. Coronary Drug Project Research Group: Practical aspects of decision making in clinical trials: The Coronary Drug Project as a case study. Controlled Clin Trials 1:363–376, 1981
8. Fleming TR and Watelet LF: Approaches to monitoring clinical trials. Natl Cancer Inst 81(3):188–193, 1989
9. DeMets D. Stopping guidelines vs. stopping rules: A practitioner's point of view. Communications in Statistics-Theory and Methods, 13(19):2395–2417, 1984
10. Jennison C and Turnbull BW: Interim analyses: the repeated confidence interval approach. J Royal Stat Soc, 51(3):305–361, 1989

Elsevier—Controlled Clinical Trials—04-07-93 14-38-57
93CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613

11. Emerson SS and Fleming TR: Interim analyses in clinical trials. Oncology, 4(3):126–133, 1990
12. Pocock SJ: Group sequential methods in the design and analysis of clinical trials. Biometrika, 64:191–199, 1977
13. O'Brien PC and Fleming TR: A multiple testing procedure for clinical trials. Biometrics, 35:549–556, 1979
14. DeMets DL: Practical aspects in data monitoring: a brief review. Stat Med 6:341–348, 1987
15. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J and Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. Br J Cancer 34:585–612, 1976
16. Haybittle JL: Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol 44:793–797, 1971
17. Simon R: "The Article Reviewed" section. Emerson/Fleming article (see Emerson SS and Fleming TR citation). Oncology, 4(3):134–136, 1990
18. Beta-Blocker Heart Attack Study Group: The Beta-Blocker Heart Attack Trial-A randomized trial of propranolol in patients with acute myocardial infarction. I. Mortality results. JAMA 247:1707–1714, 1982
19. DeMets D, Hardy R, Friedman L, Lan G: Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. Controlled Clin Trials, 5:362–372, 1984
20. Moertel CG, Fleming TR and Macdonald JS: Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. N Engl J Med 322:352–358, 1990
21. Volberding PA, Lagakos SW, Koch MA, Pettinelli C, Myers MW, Booth DK, Balfour HH, Reichman RC, Bartlett JA, Hirsch MS, Murphy RL, Hardy D, Soeiro R, Fischl MA, Bartlett JG, Merigan TC, Hyslop NE, Richman DD, Valentine FT, Corey L, and the AIDS Clinical Trials Group of the National Institute of Allergy and Infectious Diseases: Zidovudine in asymptomatic human immunodeficiency virus infection. N Engl J Med 332:941–949, 1990
22. International Chronic Granulomatous Disease Study Group: A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease. N Engl J Med 324(8):509–516, 1991
23. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators: Preliminary Report: Effect of Encainide and Flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. N Engl J Med 321(6):406–482, 1989
24. Rosner GL and Tsiatis AA: The impact that group sequential tests would have made on ECOG clinical trials. Stat Med, 8:505–516, 1989
25. Green SJ and Fleming TR: Guidelines for the reporting of clinical trials. Seminars in Oncology, 15(5):455–461, 1988
26. DeMets DL and Ware JH: Asymmetric group sequential boundaries for monitoring clinical trials. Biometrika, 69:661–663, 1982
27. Emerson SS and Fleming TR: Symmetric group sequential test designs. Biometrics, 45:905–923, 1989
28. Halperin M, Lan KKG, Ware JH, Johnson NJ and DeMets DL: An aid to data monitoring in long-term clinical trials. Controlled Clin Trials 3:311–323, 1982
29. Lan KKG, Simon R and Halperin M: Stochastically curtailed tests in long-term clinical trials. Communications in Statistics-Sequential Analysis, 1:207–219, 1982
30. Lan KKG and Wittes J: The B-value: A tool for monitoring data. Biometrics, 44:579–585, 1988
31. Spiegelhalter DJ, Freedman LS and Blackburn PR: Monitoring clinical trials: conditional or predictive power? Controlled Clin Trials 7:8–17, 1986

32. Fleming TR, Harrington DP and O'Brien PC: Designs for group sequential tests. Controlled Clin Trials, 5:348–61, 1984
33. Meier P: Statistics and medical experimentation. Biometrics 31:511–529, 1975
34. Marx JL: Drug availability is an issue for cancer patients, too. Science, 245(4916):346–457, 1989
35. DeMets DL and Gail MH: Trial of logrank tests and group sequential methods at fixed calendar times. Biometrics, 41:1039–1044, 1985 (December)
36. Lan KKG and DeMets DL: Changing frequency of interim analysis in sequential monitoring. Biometrics, 45:1017–1020, 1989
37. Lan KKG and DeMets DL: Discrete sequential boundaries for clinical trials. Biometrika, 70:659–663, 1983
38. Healy B, Campeau L, Gray R, et al: Conflict of interest guidelines for a multicenter clinical trial of treatment after coronary artery bypass graft surgery. N Engl J Med 320:949–951, 1989
39. Green SJ, Fleming TR and O'Fallon JR: Policies for study monitoring and interim reporting of results. J Clin Oncol 5:1477–1484, 1987
40. Schwarz R: Maintaining integrity and credibility in industry sponsored clinical research. Controlled Clin Trials (to appear)
41. Emerson SS and Fleming TR: Parameter estimation following group sequential hypothesis testing. Biometrika 77:875–892, 1900.
42. Jennison C and Turnbull BW: Confidence intervals for abnormal parameter following a multistage test with application to MIL-STD 105D and medical trials. Technometrics, 25:49–58, 1983
43. Tsiatis AA, Rosner GL and Mehta GR: Exact confidence intervals following a group sequential test. Biometrics, 40:797–803, 1984
44. Whitehead J: On the bias of maximum likelihood estimation following a sequential test. Biometrika, 73:573–81, 1986
45. Chang MN and O'Brien PC: Confidence intervals following group sequential tests. Controlled Clin Trials, 7:18–26, 1986
46. Kim K and DeMets DL: Confidence intervals following group sequential tests in clinical trials. Biometrics, 43:857–864, 1987
47. Chang MN: Confidence intervals for a normal mean following a group sequential test. Biometrics, 45:247–254, 1989
48. Jennison C and Turnbull BW: Repeated confidence intervals for group sequential clinical trials. Controlled Clin Trials, 5:33–45, 1984
49. Emerson SS and Banks PLC: Estimation of secondary outcomes following a group sequential trial. Technical Report, University of Arizona, 1992
50. Fleming TR: Evaluation of active control trials in AIDS. J AIDS, 3:S82–S87, 1990
51. Fleming TR: Treatment evaluation in active control studies. Cancer Treat Rep, 17(11):1061–1065,1987
52. Cancer P: Monitoring long-term clinical trials for beneficial and adverse treatment effects. Commun Statis—Theory and Methods 13:2369–2394, 1984

lsevier—Controlled Clinical Trials—04-07-93 14-38-57
3CCT—Page Design 29—File 23 Page Unit: a23 ELSEVIER CCT 613