The Early Termination of Clinical Trials: Causes, Consequences, and Control
With Special Reference to Trials in the Field of Arrhythmias and Sudden Death

Task Force of the Working Group on Arrhythmias of the European Society of Cardiology

Abstract The early termination of clinical trials, for either benefit or harm, often generates undue enthusiasm or alarm. The enhanced publicity attending early termination of a trial promotes inappropriate interpretations that are favored by the inherent difficulty of prompt and comprehensive data review. Furthermore, the process of monitoring the accumulating outcome data for early evidence of treatment benefit or harm is fraught with many statistical and methodological difficulties. This report from a task force convened by the Working Group on Arrhythmias of the European Society of Cardiology incorporates first, a series of trials terminated appropriately or inappropriately for benefit or harm and used as examples to illustrate the importance of suitable trial design and of proper stopping rules; second, a description of the committee structure of a clinical trial; third, an analysis of the general design issues; fourth, a review of the main issues in interim analysis with special reference to main strategies for reducing the rate of false-positive claims that could result from early trial termination; and finally, a series of specific recommendations concerning the design, structure, analysis, interpretation, and presentation of a clinical trial. (Circulation. 1994;89:2892-2907.)

Key Words • antiarrhythmic drugs • cardiac mortality • interim analysis

This report concerns the early termination of clinical trials of antiarrhythmic treatments within the context of concept, design, and conduct of clinical trials in cardiology. There have been several recent notable examples of trials in the cardiovascular area that have been terminated earlier than planned for either benefit or harm. The reason for stopping the trial has been soundly based in some cases but questionable in others.

The primary objective of this report is education. The intention is to help the general cardiologist and clinical cardiac electrophysiologist understand the principles of the conduct and interpretation of a clinical trial. In particular, the focus will be on the early termination of trials investigating the effects of antiarrhythmic treatment on mortality. These issues are rarely discussed in the cardiological or general medical literature and are usually confined to the pages of specialist publications. The insufficient communication between statisticians and clinicians may be partly due to the lack of a common language. Accordingly, this document has an "Appendix" with a brief glossary.

Although drug therapy and surgery have been the subject of the majority of trials in cardiology, there is now a rapid increase in trials involving procedures and devices for the treatment of cardiac arrhythmias. The recommendation of the recent "Temple Report"[1] that new devices, such as implantable cardioverter-defibrillators and pacemakers, should be submitted to randomized, controlled trials will stimulate more clinical trials. Thus, there will be sponsors and investigators relatively new to the discipline of the clinical trial.

Early and especially premature termination of a clinical trial could lead to failure to document the true efficacy or safety of an intervention. In turn, this might deprive the community of a useful intervention or it could permit an ineffective or unsafe intervention to be approved or to remain on the market. The enhanced publicity attending early termination of a trial further promotes inappropriate interpretations.

The flourishing of clinical trials has increased the likelihood that a group of investigators may embark on such a trial without full awareness of the correct methodology. Whenever this happens, the risks of an im-
TABLE 1. Reasons for Early Termination of Clinical Trials

1. Based on accumulated data from the trial:
   A. Unequivocal evidence of treatment benefit or harm
   B. Unexpected unacceptable side effects
   C. No emerging trends and no reasonable chance of demonstrating benefit
2. Based on overall progress of the trial:
   A. Failure to include enough patients at a sufficient rate
   B. Lack of compliance in a large number of patients
   C. Insufficient funding
3. Based on external information:
   A. Data from other trials establishing unequivocal treatment benefit or harm
   B. Data from clinical practice indicating unsuspected and unacceptable treatment side effects
   C. Developments superseding current treatment
   D. Treatment withdrawal from the market

proper early termination increases, largely because of the lack of an experienced and/or independent Data and Safety Monitoring Committee (DSMC), because of inadequate stopping guidelines that would prevent overreaction to a favorable or unfavorable trend, or because of pressure from the sponsor.

Early termination often generates undue enthusiasm or alarm that is intensified by the inherent difficulty of prompt comprehensive and critical review because a preliminary report is usually hastily prepared and a final report appears long after the termination date. Moreover, the process of monitoring the accumulating outcome data for early evidence of treatment benefit or harm is fraught with many statistical and methodological difficulties. Therefore, there is a considerable risk for highly publicized but misguided conclusions, particularly when the investigators and the trial sponsor have insufficient expertise in clinical trials.

The Working Group on Arrhythmias of the European Society of Cardiology assembled a task force to discuss the principles and practice of the early termination of clinical trials and to formulate recommendations.

Reasons for Early Termination of a Clinical Trial

The decision to terminate a clinical trial should be based on (1) outcome data accumulated during the trial, (2) data related to the overall progress of the trial, and (3) any external information (see Table 1).

The reasons listed in Table 1 under 1 are related to interim analyses involving monitoring of outcomes by treatment. The present report focuses largely on this aspect and on which appropriate decisions can be made.

The reasons listed under 2 are related to administrative monitoring and are briefly discussed.

The reasons listed under 3 are either self-explanatory or create an ethical dilemma related to the appropriateness of continuing the trial given the new information.

Examples of Early Termination of Clinical Trials

The complexity of data monitoring and decisions of early trial termination are best illustrated by examples. A decision to stop a trial due to a beneficial effect should require convincing evidence. However, extending a trial beyond the point of emergence of conclusive findings raises ethical issues, particularly in a placebo-controlled trial. A decision to terminate a study due to harm is often more complex. Harm, or at least the magnitude of harm, is usually unexpected. Ethically, it is argued that one could accept less conclusive evidence than for a beneficial effect. In either case, the strength of the evidence must take into account the statistical issue of "multiple looks" or repeated testing of accumulating data. A more debatable reason for trial termination is that a conclusive positive or negative trial outcome can be excluded with very high probability.

Although a decision to terminate a trial early is often judgmental, adherence to several research principles is essential in order to avoid improper trial termination. To illustrate these principles, we will present examples of trials considered properly terminated, trials judged as having been improperly stopped for a variety of reasons, and a trial with premature reporting of results.

Properly Terminated Trials

The Beta-blocker Heart Attack Trial (BHAT) was a double-blind, randomized clinical trial comparing the effectiveness of propranolol (relative to placebo) in reducing total mortality in 3837 hospital survivors of acute myocardial infarction. The trial was terminated on recommendation of the DSMC 9 months before the scheduled closing date. At the time of the decision, 7.2% of the propranolol-treated patients had died compared with 9.8% of the placebo patients, a relative mortality reduction of 26%. Many issues were considered in the decision to stop the trial ahead of schedule. These included the magnitude of the overall results, consistency of results across subgroups, clinical centers, and causes of death, and completeness of follow-up. In addition, the DSMC applied appropriate statistical methods to take into account the issue of repeated significance testing. The results were statistically significant by conventional standards at the second interim analysis in October 1979, but the trial was continued for another 2 years until the evidence met statistically prespecified criteria to justify early termination of the trial (see below).

The Cardiac Arrhythmia Suppression Trial (CAST) was a placebo-controlled study aimed at evaluating the effect of three antiarrhythmic drugs (flecainide, encainide, and moricizine) in patients with a myocardial infarction and with asymptomatic or mildly symptomatic ventricular arrhythmias. CAST was unique because it was stopped early in two steps by an independent DSMC. Less than 2 years after its initiation, part of the study was stopped because of excess mortality in patients treated with encainide and flecainide: 7.7% compared with 3.0% in the placebo group. The moricizine arm was continued but was terminated 1 year later, based on information from stochastic curtailment, a technique that calculates the probability of a reversal in the current trend in the data, and because of excess
mortality compared with placebo in the run-in phase. The incorporated monitoring boundaries, similar to but slightly less conservative in the early period than the type of boundaries proposed by O'Brien and Fleming, contributed to a proper early termination of the study. Many physicians were upset because they learned of this decision from their patients or from the press, but there is no doubt that the decision itself was correct.

Improperly Terminated Trials?

The Cardiac Arrest Study Hamburg (CASH) is a study in survivors of cardiac arrest due to documented ventricular tachycardia (VT) or ventricular fibrillation (VF) to “compare the incidence of sudden death (SD), cardiac mortality, and total mortality among four treatment groups: amiodarone, propafenone, metoprolol, and the implantable cardioverter-defibrillator.” CASH planned to enter 100 patients in each group. After inclusion of 60% of the patients, there were no differences in observed total mortality. However, the investigators observed five nonfatal VF and four of VT in patients treated with propafenone and to a lesser extent in patients treated with the two other antiarrhythmic drugs. The combined incidence of total mortality, of resuscitated VF, and of VT in the propafenone group (33%) was compared with the incidence of total mortality in the implantable cardioverter-defibrillator (ICD) group (12%). The probability value resulting from this comparison was <.05, which was the reason given for terminating the propafenone arm of the trial. The investigators claimed that propafenone could not be recommended in this high-risk population. However, this claim cannot be substantiated on the basis of this trial for the following reasons: (1) the investigators compared dissimilar combinations of events between the two groups, and the choice of the combined end point a priori favored the ICD treatment; (2) no statistical corrections were applied for continuous monitoring (multiple looks at the data). With continuous data monitoring, the probability of reaching a level of P<.05 (which means a probability <5%) by chance alone is actually much higher than 5%. This implies a high likelihood of a false-positive result; (3) the trial encompassed three pharmacological groups. By selecting the treatment group with the highest incidence of nonfatal VF and VT and comparing it with the ICD group, the risk of a false-positive result favoring the ICD group was substantially increased. Appropriate corrections for multiple comparisons should have been applied.

Continuous interim analysis of events during a trial may lead to improper early termination. In the extreme case, when a trial is monitored every time a death occurs, the likelihood of obtaining a statistically significant probability value is quite high, when in reality there is no difference. Such a "nonadjusted" (and hence incorrect) probability value may evoke an emotional temptation to stop the trial prematurely. A sequential boundary for continuous data monitoring can be used to control the rate of this so-called type I error at the desired level. Very often, this is not done, as shown by the following two examples.

The first example is the Timolol, Encaidine, Sotalol Trial (TEST). In this trial, the United States Food and Drug Administration (FDA) was provided with on-line access to major outcome data. This unusual arrangement contributed to early termination after only 13 deaths were reported. This could have been avoided had the appointed independent committee been solely responsible for the monitoring of the outcome data in accordance with the principles presented in this document.

The second example is the diltiazem-bepridil study of patients with angina pectoris (E. Prystowsky, personal communication). After enrollment of approximately 80 patients per study arm, there were five deaths in the bepridil group compared with none among the patients randomized to receive diltiazem. The sponsor decided to terminate the trial despite advice to continue. The difference in mortality rates did not reach statistical significance. Had a well-rationed protocol been established for monitoring, the trial might have continued and provided more reliable and informative results.

Prematurely Reported Trials

The Anturane Reinfarction Trial (ART) was a randomized, double-blind, placebo-controlled trial of sulfipyrazone in survivors of a recent acute myocardial infarction. ART was organized, conducted, and analyzed by the sponsor, CIBA-GEIGY Corporation, USA. A total of 1629 patients were enrolled between 25 and 35 days after the acute event and were followed for a minimum of 1 year (average, 16 months). The ART was fraught with many violations of established principles for the design and analysis of clinical trials. The primary outcome of cardiac mortality was changed to sudden death between the two trial reports. A large number of the randomized patients (N=71) were excluded from the final report as being ineligible or "nonanalyzable." Many of these decisions were made post hoc, after the patients had been treated for extended periods and died. The exclusion of 10 deaths in the sulfipyrazone group versus 4 deaths in the placebo group influenced the reported significance level in favor of sulfipyrazone. The early reporting of the mortality findings in 1978 after an average follow-up of 8.4 months received considerable attention. The trial was criticized for not having "stopping rules" and for lack of active involvement of an independent DSMC. Due to these limitations of ART, the FDA disapproved the application for a new indication for sulfipyrazone. Moreover, FDA officials took the unusual step of publishing their critique of the trial. A review of the data on cause-specific mortality by an independent committee further discredited the reported findings regarding sudden deaths.

Committee Structure of a Clinical Trial, With Special Reference to Its Role in Early Termination

Although many committees conduct important work for a clinical trial, we discuss here the committee structure involved with the decision to stop a trial early. When organizing a clinical trial, it is important to identify who has responsibility for (1) monitoring and summarizing data that form the basis for the decision to stop a trial, (2) reviewing the data summaries and recommending stopping, and (3) making the decision to stop the trial. Multicenter clinical trials that have mortality or a surrogate for mortality for the primary end point should
be structured to avoid appearance of conflict of interest for the sponsor. The sponsoring agency should utilize an independent DSMC to monitor the trial and consider its recommendations.

**A Clinical Trial Sponsored by the National Heart, Lung, and Blood Institute**

The committee structure of randomized clinical trials initiated and sponsored by the National Heart, Lung, and Blood Institute (NHLBI) is shown in Fig 1. Each clinical center has a principal investigator and research coordinator who are responsible for recruiting patients and for submitting data to the Data Coordinating Center (DCC) on a regular schedule. The DCC is responsible for entering, correcting, and monitoring the data. The DCC tabulates data related to recruitment, data quality and timeliness, adverse events, and death. One person, usually a statistician, tabulates the deaths by treatment group assignment on a regular basis.

Administrative and outcome data are reviewed on a prespecified schedule by the DSMC, an independent group of clinical trialists, statisticians, clinical cardiologists, and ethicists appointed by the NHLBI to review data from the trial and act on behalf of the patients who enroll.

Meetings of the DSMC often are attended by several nonvoting members: the NHLBI project officer and an NHLBI statistician, the trial statistician and a few other staff members from the DCC, and the chairman of the steering committee, who is one of the clinical center principal investigators appointed by the NHLBI. The chairman’s role at meetings of the DSMC is to help the DSMC identify problems early and seek corrective action. When interim analyses are discussed, the chairman of the steering committee and other advisory personnel usually leave the meeting. Representatives of collaborating commercial units, eg, a pharmaceutical company, do not attend meetings of the DSMC.

The DSMC meets at regular intervals to discuss interim summaries of the recruitment, the timeliness and accuracy of data collection, the number and nature of adverse events, and the mortality rates. Under special circumstances, an early meeting of the DSMC may be called. Various strategies with respect to blinding are used to present outcome data to the DSMC. They should be presented in a way that permits the DSMC to evaluate fully the risks and benefits of the treatment group compared with the control group. If outcome data are submitted to the DSMC in coded form, the DSMC can request that the code be broken whenever a concern arises, for example, when the data show that the treatment group difference is approaching a stopping boundary. If the possibility of a serious adverse effect is suspected, the chairman of the DSMC should review adverse effects, including mortality data, between regular committee meetings.

If the DSMC recommends that the trial be stopped, the Director of the NHLBI is responsible for the decision to stop the trial and notify collaborating units and the participants. In case of stopping for harmful treatment effects, the commercial sponsor and the US Food and Drug Administration or similar agency must be notified immediately.

**Investigator-Initiated Clinical Trials Funded by the NHLBI**

The structure of a randomized clinical trial that is initiated by a group of investigators and funded by the NHLBI is similar to that for an NHLBI-initiated trial. The first trial personnel to hear the recommendation to stop a trial will be the DCC statistician who attends the DSMC meeting and the principal investigator. Although there is some variation, most trials have an executive group of investigators appointed by the principal investigator who reviews a DSMC recommendation to stop the trial early. The final decision to stop the trial usually rests with the executive committee or with the principal investigator alone.

**Randomized Clinical Trials With a Commercial Sponsor**

A commercial sponsor inevitably has a vested interest in trials providing results that could support a new claim for a product. The credibility of a trial is enhanced if the results are seen to be free of influence from the sponsor. Credibility is greatest when a clinical trial is conducted, analyzed, and reported independently of the sponsor.

Several structural models have been used to conduct randomized clinical trials that are funded by a commercial sponsor. One structure, common in international studies, is shown in Fig 2.

Two interesting differences exist between the NHLBI model and the model recommended for commercially sponsored studies. First, in commercially sponsored studies, a shared arrangement for trial design, data management, safety monitoring, and data analysis is often used. Second, the functions served by a single DSMC in an NHLBI trial often are divided between two committees in trials with a commercial sponsor: the DSMC, a group that reviews the interim analyses and considers early termination of the trial, and the Policy
often have been used for substantially contributing and obtaining personnel for Board, a group responsible for the administrative matters for the trial that has the authority to stop the trial.

Commercial sponsors, eg, pharmaceutical companies, often have highly developed and effective systems for obtaining and checking data, and these resources can contribute substantially to the success of clinical trials. It is appropriate that a sponsor’s data management resources be used for a clinical trial provided that their personnel do not have access to the randomization code or outcome data and they cannot become unblinded. In the structure shown in Fig 2, responsibility for data management is shared by the sponsoring company and an independent clinical trials unit in a way that preserves the integrity of the trial.

In years past, it was common for clinical trials to be conducted entirely by the sponsoring company, including randomization and management of the outcome data. The review process was even conducted by statisticians who were employed by the sponsor. Statistical summaries were reviewed by scientists who worked for the sponsor and even by some of the sponsor’s management team. We believe that this degree of sponsor involvement in a clinical trial severely compromises the credibility of a trial.

For safety monitoring in the shared-data management arrangement, the commercial sponsor’s data management unit should provide timely administrative data, data on baseline patient characteristics, etc., to the external clinical trials unit so it can prepare regular reports for the DSMC.

After each DSMC meeting, the chairman should write a letter to the appropriate regulatory agency to satisfy the sponsor’s regulatory requirements to report adverse experiences. This arrangement obviates the need for the commercial sponsor to have access to data that link the treatment assignments to the primary outcome. Regulatory agencies appreciate the desirability of avoiding conflict of interest and are willing to cooperate with efforts to keep data monitoring during the trial and the analysis of the data after the trial separate from the commercial sponsor’s activities. However, regulatory affairs are complex, and individualized arrangements must be made carefully for each trial.

In the model shown in Fig 2, a recommendation by the DSMC to stop a clinical trial early may be submitted to the Policy Board, an independent group of experts who attend to the administrative matters of the clinical trial and are entrusted with the authority to stop the trial. Alternatively, an executive committee of the investigators may be the group to receive the DSMC recommendations and to make the decision to stop the trial. When the external Policy Board or the Executive Committee of the investigators decides to accept the recommendation of the DSMC to terminate the trial early, the commercial sponsor and the principal investigator are notified. The principal investigator, in turn, notifies the investigators in the clinical centers.

When the clinical trial ends, there are additional benefits to the model shown in Fig 2. The external clinical trials unit should analyze the data needed by the investigators for the primary and most important secondary publications. The sponsor’s statistical unit performs the data summaries and analyses to satisfy regulatory requirements and new claims for the efficacy and safety of the product that was evaluated in the clinical trial. The sponsor gains substantial credibility and promotes acceptance of the trial findings by avoiding any possible appearance of conflict of interest.

**General Design Issues**

The basic requirement for any clinical trial is that a well-defined primary question should be stated. Otherwise, neither interim nor final analysis can be performed rigorously. Furthermore, the protocol needs to specify clearly which subjects are to be included in the analyses and which, if any, subgroups and secondary outcome variables are to be considered. Post hoc–defined subgroups and secondary variables are not reliable indicators of treatment effect because some may show trends by chance alone.

**Definition of Outcomes**

The lack of a precise definition of outcomes has been a major deficiency in a number of trials such as the Anturan Reinfarction Trial (ART).\(^{12-14}\) a lack that has been particularly relevant in trials of antiarrhythmic therapy. It is absolutely essential that outcomes such as sudden death be defined in the original protocol of a trial. Unfortunately, there is no international agreement as to how to categorize sudden death, death while sleeping, arrhythmic death, and death due to congestive heart failure, etc.\(^{15}\) This makes meaningful comparison between studies virtually impossible. When “soft” end points such as myocardial infarction or sudden death are important outcomes, these should be categorized by an independent “end point” or “validation” committee blinded to the treatment allocation.

This Task Force strongly recommends that an international committee be formed in the near future to address this issue.
Multiplicity of Outcomes

Multiple outcome variables increase the chances of claiming a benefit or harm by chance alone. Clinical trials usually define a primary outcome variable and relegate other variables to a secondary status. The primary outcome variable determines the eventual result of the trial. It is used as the decisive variable in the process of interim analyses. Secondary outcome variables constitute supportive evidence in interim analyses as well as in any final analyses.

Multiple Treatment Groups

When there are more than two treatment groups, there is the problem of multiple treatment comparisons, which implies that, for example, when mortality is compared in the two groups with the most extreme results, there is a much higher probability of obtaining significant results by chance alone than the usual 5%. There are standard statistical methods to make adjustments so that a proper significance level is maintained. However, when the focus is a priori on one particular comparison (eg, drug A versus placebo), these adjustments are usually not necessary.

Intention to Treat

Differences of outcome in clinical trials can only be regarded as real if the following assumptions can be validated: (1) that the treatment strategies have been applied to comparable groups of patients and (2) that the outcomes assessment has not been influenced by knowledge of the treatment strategy.

Blinding of both trial personnel and the patient to the treatment strategy (double blind) is one way to ensure the validity of the second assumption. Another is to choose an outcome that is unequivocal, eg, death. In the case of an equivocal outcome involving the mechanism of death such as sudden or arrhythmic death, it is imperative that there is blinding of assessment from therapy. This poses great difficulties for obvious treatments such as the implantable defibrillator and can jeopardize the credibility of a trial when not achieved, as in the ART,12,13 in which the classification of sudden death was challenged.14

The best possible way to obtain comparable groups of patients is to randomize the group assignment. Even randomization does not ensure comparable groups, but it does ensure that for large numbers of patients, any difference between groups is random and small. Removal of patients before the attainment of the end point can jeopardize the comparability of the groups.

Unfortunately, no trial is perfect in maintaining the intended treatment throughout. For example, a patient may die, cross over to an alternative treatment strategy, move, be assigned but not receive a specified treatment strategy, or otherwise make initiation and/or continuation of treatment impossible. Patients who refuse to comply with the treatment strategy almost certainly are biased subgroups, and the bias probably depends on the treatment strategy. For example, if one strategy takes longer to implement, more deaths might be expected to occur before that strategy becomes effective, and thus the patients actually receiving the treatments will differ in risk. Patients in the placebo arm of the Coronary Drug Project who were noncompliant were found to be at significantly higher risk than compliant patients16 (Table 2). This finding has been replicated in other studies.17 There is no reason to believe that patients noncompliant to an active drug will be similar to patients noncompliant to placebo because the former probably are reacting to drug effects, whereas the latter assuredly are not. Similarly, there is no reason that patients withdrawn by protocol or who cross over between treatment strategies will form comparable subgroups, eg, patients who cross over from drug to surgical treatment are likely to have a more severe disease. Thus, the only valid comparison that can be made must be between groups as defined by randomization, ie, according to intention to treat and not according to actual treatment (so-called efficacy analysis). Intention to treat implies that all events are counted until the end of the trial irrespective of whether or not the patient is maintained on the assigned treatment strategy. In survival analyses, censoring is done only because the study has ended or death has occurred.

The temptation to perform an efficacy analysis becomes stronger, the greater the difference in the number of patients who are assigned to treatment relative to the number actually receiving treatment. For example, in the Coronary Drug Project, only 67% of patients were ≥80% compliant with clofibrate treatment. A typical efficacy analysis might have compared the compliant clofibrate patients (15% mortality) with the placebo and noncompliant clofibrate patients (>20% mortality) and reported a significant beneficial effect for clofibrate (Table 2). The solution is not to plan to perform efficacy analyses as the primary analyses; rather, the design should require adequate numbers of patients to achieve the objective despite the dilution of treatment effect by this phenomenon. Great care must be taken during the design phase to reduce the inconsistencies between intention and treatment. For example, a “run-in” phase may identify most noncompliers, and these should not be enrolled into the randomized comparison trial. Deaths between randomization and initiation of treatment can be reduced by not randomizing until the last possible moment.

Efficacy analyses may provide some insight into the interpretation of trials with poor compliance. It will not rescue the results of such a trial but may help address compliance issues in the ongoing trial or in the design of future trials. Such analyses must be done with appropriate caution because in many cases, one cannot predict the direction of potential biases.

A consequence of the intention-to-treat principle is that one must accept that a trial often compares strategies rather than particular interventions. For example, one may wish to compare the ICD to amiodarone; in reality, one will compare the strategies of initial ICD versus initial amiodarone treatments. This approach

<table>
<thead>
<tr>
<th>Drug Compliance</th>
<th>Overall</th>
<th>≥80%</th>
<th>&lt;80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clofibrate</td>
<td>18.2</td>
<td>15.0</td>
<td>24.6</td>
</tr>
<tr>
<td>Placebo</td>
<td>19.4</td>
<td>15.1</td>
<td>28.2</td>
</tr>
</tbody>
</table>
actually mirrors the clinical approach to therapeutic decision making.

**Subgroup Analysis**

Subgroup analyses are often used to examine consistency of results for positive or negative results without regard to statistical significance. However, for trials with neither negative nor positive trends, such analyses may identify subgroups associated with treatment benefit or harm in which the association will generally be spurious.\(^{18}\) If external information identifies a subgroup of interest that can be addressed by a trial, then legitimate hypotheses can be tested. If the subgroup hypothesis is generated by data exploration from the study itself, a great deal of caution must be exercised before conclusions are drawn and decisions made.

Analysis by subset may suggest that there are certain subgroups in which there appears to be overall benefit or harm, eg, the elderly or those with impaired left ventricular function, but without reaching significance. Under these circumstances, it may be extremely difficult to decide whether to continue enrolling a subset or to stop the whole trial. Whenever possible, subgroups should be defined external to the trial and a priori.

**Issues in Interim Analysis**

**Ethical and Clinical Issues**

Is the primary responsibility of an investigator undertaking a clinical trial to the individual patients in that trial or to the community at large?\(^{19-21}\)

The conduct of a clinical trial depends on the consent of individual patients to participate in it. The investigator is morally obliged to conform with the agreement reached with the patient. There is normally a written or tacit understanding that the trial is being undertaken because it is uncertain whether the treatment being tested is superior to "standard" treatment for that condition. If the new treatment is superior, there is a significant chance that, by agreeing to take part, a participant would benefit. If unequivocal evidence of benefit emerges in the treatment group, it would be ethically indefensible to continue to administer placebo unless the patient has agreed to such a policy for the common good. There is also a possibility of harm, but it is understood that if unequivocal evidence of such harm emerged during the trial, the patient's participation should be terminated without delay. However, evidence of effectiveness or harm is based on probability, not on absolute proof. Often, results cannot be regarded as unequivocal, and it is then appropriate to continue the trial to obtain more data. It may not be easy to make decisions based on these principles. For example: (1) The treatment under study may not be available, so that to stop the study would be to deprive the patients destined to receive the active therapy of its benefits. This was addressed in the AIMS study,\(^{22}\) which showed a 50% reduction in mortality from anistreplase by stopping the study when the statistical guideline had been reached and giving the active agent to the patients who would have been recruited into either the active or placebo groups. (2) The treatment under study may be available but not in common use, so that the patients yet to be recruited might not receive the treatment. This was an argument for continuing ISIS-2 despite clear benefit in those treated with streptokinase within 4 hours of onset of symptoms.\(^{23}\)

**Administrative Monitoring**

Designing a clinical trial is always a challenge. A trial is based on several initial assumptions. It is not unusual that one or more of these assumptions are incorrect and are only discovered after the trial has started. Close monitoring of the progress of the trial and quality control of the incoming data may identify invalid assumptions used in the design, such as enrollment of the wrong population or risk group, unrealistic recruitment goals, difficulty with outcome measurement, and problems in treatment compliance or data collection. If identified early, design modifications might be made to correct serious flaws. For example, if recruitment is slow or low-risk subjects are being enrolled, entry criteria may have to be changed. Perhaps a dose modification or strategy modification is required to enhance compliance. Frequency or timing of measurements may be altered to accommodate patients better or to enable measurement of the end point. Constant vigilance of the data quality is necessary, and early data analyses may identify problems that can be corrected.

These design and logistic analyses are referred to as administrative monitoring. If conducted early without treatment comparisons, no formal structure is required, although planning for the common problems noted above is advisable. However, if administrative monitoring is conducted later in the trial, when data on primary and secondary outcomes are emerging, special care must be taken to ensure that treatment comparisons are not involved. Any analysis using outcome data must be subject to procedures described in the interim analyses section. Usually, administrative monitoring is most helpful in the early stages, when there is still time for meaningful design corrections.\(^{24}\)

**Decision Factors and Early Termination**

The decision to terminate early for benefit or harm or lack thereof is quite complex,\(^{19-21}\) and many factors must be considered before a study is stopped (Table 1). Major considerations are baseline comparability of subjects for important risk factors and consistency of data, internally across primary and secondary outcomes and subgroups and externally with other studies. Of course, the benefit of treatment compared with risk must be assessed, taking compliance into consideration.

**Repeated Significance Testing**

It is well known that repeating statistical tests on a single outcome over time will increase the likelihood of finding spurious treatment differences.\(^{25}\) An excellent illustration based on actual data is provided by the Coronary Drug Project (CDP), as shown in Fig 3.\(^{19}\) One treatment, clofibrate, showed no mortality difference compared with placebo after 100 months of follow-up. However, if the clofibrate-placebo--standardized treatment differences were repeatedly computed as data accumulated over time, the standard \(Z\) test statistic would have approached a critical value of 2.0 (nominal unadjusted \(P= .05\)) on four occasions. However, the CDP investigators were aware of the problem posed by repeated testing and did not terminate this arm of the study. They developed several techniques to adjust for
repeated testing, which were early versions of two of the monitoring methods often used now, group sequential methods and stochastic curtailment.

The extent of false-positive claims, or the type I error rate, has been well quantified. If a two-sided critical value of ±1.96 for the Z test, corresponding to a nominal 5% significance level, is used for each test, the type I or false-positive error for a single test will be .05, as designed. However, the error rate is .08 for 2 interim analyses, .14 for 5 interim analyses, and .19 for 10 interim analyses. Although there is no exact rule on how large the type I error rate should be, even a 1-in-10 chance of a false-positive claim is too large for pivotal clinical trials.

There are four basic strategies for reducing the rate of false-positive claims that could result from early trial termination: (1) group sequential methods that use only a few monitoring inspections of accumulating data in a planned fixed sample size trial; (2) stochastic curtailment, which projects results at the end of a trial from emerging data; (3) Bayesian analyses, which incorporate prior and emerging data into the decision process; and (4) full sequential trial designs with prespecified single- or double-triangular stopping boundaries. Each strategy attempts to be conservative in reacting to trends before claiming a conclusive treatment effect while preserving sensitivity to real treatment differences. We will discuss only the first two of these strategies in the sections that follow.

Group Sequential Methods

In the classic fixed sample design, all of the allowed type I error rate or α level is used only once, at the final analysis. In contrast, group sequential methods control the α level by allocating some of the prespecified α level to each interim analysis.

In the design of a survival study, an estimate of the treatment effect is made. Then, the total number of events (amount of information) required to reach a specific power or likelihood that the trial will detect such a treatment effect is calculated. When data are monitored at some time before the end of the study, the ratio

\[
\frac{\text{number of events observed at monitoring}}{\text{total number of anticipated events}}
\]

is called the information fraction at that point in the study. The value for t must be between 0 and 1. The Z value of a test statistic evaluated at that point will be denoted by \(Z_t\). A group sequential monitoring procedure can be characterized by a spending function \(\alpha(t)\), which states how the prespecified \(\alpha\) is allocated at each interim analysis relative to the fraction of information or events. In Fig 5, a spending function \(\alpha(t)\) is defined on the unit interval where the value of \(\alpha(t)\) increases from 0 at \(t=0\) (no events) to \(\alpha\) at \(t=1\) (total anticipated events).

When data are monitored for the first time at information fraction \(t_1\), \(\alpha(t_1)\) is the amount of \(\alpha\) allowed to be spent, and this amount determines a boundary or critical value for the test statistic Z. If significance is not reached at the first data monitoring, and additional data are monitored at a second time, with information fraction \(t_2\), additional \(\alpha\) (denoted \(\Delta \alpha\) in Fig 4) is spent. This difference determines a boundary or critical value for Z at \(t_2\). Future boundary values, if needed, will be determined in a similar manner.

The choice of the spending function will determine how interim monitoring affects early termination and should be made to reflect the ethical and pragmatic concerns of the trial. Consider two spending functions that allocate an \(\alpha\) level of .05 over five analyses at times corresponding to information fractions of .20, .40, .60, .80, and 1.0. The same prespecified .05 \(\alpha\) level (sum of \(\alpha\) values \(t_1\) to \(t_5\)) is spent in both these group sequential monitoring plans.
Table 3. Critical Values for Test Statistics

<table>
<thead>
<tr>
<th>Monitoring time</th>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>( t_3 )</th>
<th>( t_4 )</th>
<th>( t_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information fraction</td>
<td>.20</td>
<td>.40</td>
<td>.60</td>
<td>.80</td>
<td>1.0</td>
</tr>
<tr>
<td>Plan A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha ) Level</td>
<td>.00001</td>
<td>.00123</td>
<td>.00774</td>
<td>.01668</td>
<td>.02439</td>
</tr>
<tr>
<td>( Z ) critical value</td>
<td>4.56</td>
<td>3.23</td>
<td>2.63</td>
<td>2.28</td>
<td>2.04</td>
</tr>
<tr>
<td>Plan B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha ) Level</td>
<td>.01595</td>
<td>.01180</td>
<td>.00909</td>
<td>.00736</td>
<td>.00619</td>
</tr>
<tr>
<td>( Z ) critical value</td>
<td>2.41</td>
<td>2.41</td>
<td>2.41</td>
<td>2.41</td>
<td>2.41</td>
</tr>
</tbody>
</table>

The first plan (A) spends its \( \alpha \) value of .05 slowly during the early part of the trial, as shown in Table 3 and Fig 5 (top). The boundaries or critical values for the test statistic are shown in Table 3 and Fig 5 (bottom). Notice that in plan A, the first analysis at \( t_1 \) corresponds to a large initial critical \( Z \) value, which decreases with each consecutive interim analysis and is only slightly larger (2.04) than a nominal .05 critical value for \( Z \) of 1.96 at the end.

The second plan (B) spends the available prespecified \( \alpha \) value of .05 faster at the beginning than at the later analyses as shown in Table 3 and Fig 5 (top). The boundaries or critical values produced by this more aggressive spending function are nearly constant throughout. Here, note that the final critical \( Z \) value, 2.41, is larger than the final value, 2.04 in plan A. This implies that a larger test statistic will be required to claim significance at the .05 level in the final data analysis. The spending function chosen will affect the power of the trial to detect a certain alternative. With the first spending function, the last critical value (2.04) is only slightly larger than the nominal (1.96) that would apply if no interim monitoring was performed, implying that the conventional sample size calculations will be approximately correct. For the second spending function, the last critical value (2.41) is substantially larger than the nominal (1.96). To achieve the same level of power, the sample size in plan B must be increased over plan A. However, plan B allows for a greater chance of earlier termination.

There are many \( \alpha \) spending functions. However, the two that are presented correspond to well-known and widely used group sequential procedures. Plan A corresponds to an O'Brien-Fleming-type boundary, and plan B corresponds to the method described by Pocock. A third plan proposed by Haybittle spends very little \( \alpha \) at each analysis (eg, .001), such that most of the \( \alpha \) is still preserved for the scheduled final analysis.

The advantage of the \( \alpha \) spending function approach is that the exact information times do not have to be known or specified in advance, and the exact number of interim analyses does not have to be prespecified. Furthermore, the frequency of interim analyses can be changed during the trial. Even if additional interim analysis is motivated by strong emerging trends, the impact of changing frequency on the overall \( \alpha \) level is nearly negligible. While this approach allows for a great deal of flexibility, the particular spending function must be specified in advance in the protocol, and no change of spending functions is allowed during the trial.

With these basic requirements, an \( \alpha \) spending function approach can be used for most interim analyses. This includes comparisons of means, proportions, survival curves using log rank or other generalized rank tests, and comparison of rates of change based on repeated measures. The interim analyses test statistics must be prespecified and should be based on the same test statistics envisioned for the final analysis.

In addition to hypothesis testing, confidence intervals provide useful information about the precision of the estimate of treatment effect. If the confidence interval
does not include zero difference, the two treatments are not equal. However, confidence intervals also can be used to rule out other differences. For example, if a new treatment that is less costly or less toxic than the standard of care were less effective by 20% or more, the advantages might not outweigh the lower effectiveness. In the framework of interim analyses, "repeated confidence intervals" can be constructed by taking the estimate of treatment benefit plus or minus a coefficient multiplied by the standard error of the estimate. The coefficient is the group sequential boundary or critical value. For example, if an O'Brien-Fleming-type plan (Table 3) is used, the 95% repeated confidence interval at the third analysis would be the estimate difference ±2.63 SE.

This repeated confidence interval approach for clinical trials has been encouraged by Meier and developed largely by Jennison and Turnbull for group sequential clinical trials.

The O'Brien-Fleming type of sequential spending plan applied to a log rank test was used in the BHAT. As shown in Fig 6, the early BHAT results showed a favorable trend but did not cross the prespecified boundaries. The trial was designed for a minimum of 3 years of follow-up. With almost a year to go, the test statistics for the mortality curves (Fig 7) crossed the boundary, indicating a significant benefit for propranolol. However, before terminating the trial, all the termination issues described in Table 1 were carefully considered. None demanded that the study be continued, although how long propranolol should be given remained unanswered. Conservative monitoring boundaries allowed the trial to progress until follow-up was reasonably mature; yet, patients who were enrolled in the study as well those who were not were allowed the benefit of this drug a year earlier.

A second example is provided by the CAST study. As described earlier, CAST was a trial evaluating three antiarrhythmic drugs relative to placebo, with mortality and sudden death as outcomes. A different spending function was used, which spent \( \alpha \) at a linear rate, very slowly during the first 80% of the information fraction and somewhat faster during the last 20%. The boundaries and the CAST results are shown in Fig 8. Initially, CAST was designed as a one-sided .05 test of treatment benefit. The DSMC recommended a .025 significance level for benefit and an "advisory" .025 level lower boundary for harm, symmetric with the upper boundary. This was advisory in the sense that the DSMC might not want to require that same standard for a negative trend but if it should occur, the results would be deemed significant. As shown, the first CAST interim analysis suggested a treatment difference with only a small number of events. The DSMC remained blinded to whether drug or placebo was superior. A number of factors were considered. The number of events was small, representing only 5% of the total expected, and the follow-up period was short. With a small number of subjects entered, the randomization might not yet have achieved balance in all important risk factors. Perhaps most important was the knowledge that the drugs being evaluated were already widely used and any early termination would...
need to be convincing that the harmful trend was not due to chance.

The DSMC recommended continuation. Between the first and second DSMC meetings, the statistical center alerted the DSMC that the previously observed trends were continuing in the same direction. An emergency DSMC conference call was held. Members were told the treatment assignment and the observed number of events. The data were alarming, and the DSMC wanted to be sure that all factors were considered to rule out any possible errors in data collection and interpretation. A large series of analyses were requested, and the DSMC met as soon as all the tests could be completed, approximately 4 weeks later. When the DSMC met, the updated analyses confirmed that sudden deaths were 33 to 9, with total mortality being 56 to 22, both in favor of placebo. The conservative monitoring boundaries were crossed, with less than 20% of the information fraction observed.

From a monitoring viewpoint, procedures such as α spending functions must be in place at the beginning, and two-sided boundaries are almost always required. In addition, timeliness of data collection is absolutely essential in trials with mortality or irreversible morbidity as measures of efficacy or safety. A DSMC cannot make a fully informed decision as to the risks and benefits of treatment if data are several months old (eg, more than 2 to 3 months). Emerging trends on outdated results may diminish or become more extreme when results are updated. In such instances of tardy data, decisions made by the monitoring committee may not be appropriate and could possibly be avoided with more current data. For safety issues, delay in reporting results may cause more patients to be at further risk of the event. For treatment benefit, tardiness in reporting may delay current and future patients from obtaining access to the treatment. In either case, the most recent information possible is necessary for the decision to continue or terminate the trial.

Stochastic Curtailment

Stochastic curtailment or conditional power is a second major method for controlling false-positive errors or α levels leading to early termination.\(^{[41,42]}\) The probability of rejecting or accepting the null hypothesis of “no treatment difference” at the end of the trial is calculated from emerging data at some information fraction t. If the trend is positive, the probability of rejecting the null hypothesis at the end of the study may be calculated, assuming a reasonably conservative trend for data for the rest of the study. If the emerging data show a negative trend, it is similarly possible to calculate the probability of rejecting the null hypothesis at the end of the trial and claim a harmful effect. If those probabilities are high, .95 or larger, early termination of the trial may be considered, knowing that the results are fairly certain. Such was the case for the BHAT.\(^{[3]}\)

However, the most useful application of this method is to assess the futility (no difference) of continuing the trial. It is possible to calculate for an emerging negative trend the probability of recovering and still rejecting the null hypothesis, claiming a treatment benefit, assuming a range of alternatives for hypothesized treatment effect. Stochastic curtailment, or alternatively repeated confidence intervals, allows discouraging trends emerging from a clinical trial to be recognized and the chances of showing a positive treatment effect of clinical importance to be effectively ruled out. Stochastic curtailment often can be used as a complementary procedure, and it is not incompatible with group sequential methods and repeated confidence intervals.

In the situation in which there is little to no chance of obtaining a significant difference at the end of the study, even with the most optimistic proposed alternative of treatment benefit, the DSMC should determine whether there is any reason to continue the trial. In CAST, the chance of recovering from such a strong negative trend was small, even assuming that the treatments were more beneficial than originally assumed. Even in trials with less dramatic negative trends, stochastic curtailment has been useful in identifying lost causes. However, whether a negative trend should be followed until it reaches significance is a difficult decision, one that a DSMC must consider carefully in the context of the disease and the treatment.

With this method, early termination will inflate the overall type I error rates. However, if this approach is applied conservatively, the inflation is very small.\(^{[3,41]}\)

Asymmetric vs Symmetric Monitoring Boundaries

Perhaps the most difficult and agonizing situation for a DSMC is how to deal with emerging negative trends. Three cases are of particular interest. Symmetric boundaries are appropriate to evaluate whether treatment A or B, both in clinical use, is better. If a new therapy is being compared with a placebo, two possibilities must be considered. If the new therapy is not widely in use, then a negative trend with little chance of showing a benefit might be chance to terminate the trial or even dismiss the drug as not being of further interest. In this case, a less conservative boundary for harm might be more appropriate.\(^{[43,44]}\) However, if one is studying a drug already approved or widely used, based on surrogates for efficacy, the same degree of evidence of harm as of benefit may be required to discourage continued use.

Both upper and lower boundaries should be decided in advance before the monitoring process begins. The PROFILE study, which was terminated because of harmful effects of rosequinian, used symmetric boundaries because the therapy was already approved and in use.\(^{[45]}\) The same was true for the PROMISE trial, which studied milrinone.\(^{[46]}\) However, CAST II,\(^{[5]}\) which was a continuation of CAST\(^{[4]}\) for the third nonapproved antiarrhythmic drug, changed to a less conservative boundary for harm because of the adverse experience of the first phase, thus lowering the threshold for acceptance of a negative result.

Stopping Procedure

It is critical to design a trial with sufficient care to avoid potential pitfalls that may attend stopping the trial. It is equally critical that guidelines be developed for stopping a trial. Appropriate methods of handling early termination of the trial should be carefully detailed before its outset.

The decision to stop a trial early may be made by the manufacturer, sponsor, regulatory authority, or investigator, because each has important responsibilities. These may be commercial, regulatory, scientific, or...
ethical. However, there are vested interests in continuing a trial to achieve a positive result that may inappropriately hasten or delay the completion of a trial. This is self-evident in the case of drug or device manufacturers, but sponsors dependent on government funding also must convince politicians of the value of their work. In this regard, nonprofit organizations are not immune from criticism and may invoke hyperbole to enhance their image. Finally, the reputation and funding of an investigator may depend on the “successful” outcome of the trial.

It is therefore recommended that the decision to terminate a trial for ethical or scientific reasons should rest primarily with the appropriate trial committee that is formally assigned this task before the start of the trial. The DSMC should review the data according to the protocol and make their recommendation to discontinue the trial based on trial data, external information, or both.47

Although the tradition of medicine and the policies of regulatory agencies may attach more significance to harm than to benefit, these divergent outcomes merit equivalent objective consideration (given equivalent differences in mortality). The use of asymmetric boundaries for stopping guidelines gives more weight to harm. Once a trial is terminated, results in favor of harm or benefit deserve equally urgent communication. However, in the case of drugs that are not approved, delay in the regulatory process reduces the need for rapid communication.

When it has been decided that a trial must terminate early, the urgency of the stopping process should depend on the reason for stopping the trial. For example, a significant difference in mortality or serious morbidity must prompt a quick response, whereas less serious end points, such as non–life-threatening drug intolerance, can initiate a less rapid process. Urgent action is not required when a trial is discontinued because of the futility of demonstrating benefit or harm.

When a trial is terminated early, the investigators should be informed first so that the trial may be discontinued quickly. However, other physicians may need to receive information almost as rapidly because their patients may be receiving the intervention. These physicians should be informed by the manufacturer or sponsor. The regulatory authority also must be informed without delay. When the intervention is already approved or in widespread use, the sponsor or manufacturer as well as the regulatory authority should be informed, and the responsibility for rapid dissemination of the information lies with the manufacturer. In the case of interventions from multiple manufacturers, for example, generic drugs, the responsibility for communicating the information must rest with the regulatory agency.

The practice of hasty presentation of the results of trials shortly after their scheduled completion, often with the considerable pressure of a deadline and without any form of prior publication allowing objective scrutiny, can lead to misunderstandings and confusion, mistaken conclusions, and ill-considered recommendations. While commonplace, it is not advisable unless there is opportunity for discussion or independent commentary. On the other hand, an early forum is encouraged for trials with early termination because of harm or unexpectedly significant benefit related to survival or serious morbidity. Ultimately, the fundamental data from all trials merit consideration for publication in an appropriate peer-reviewed journal, regardless of the trial outcome. Such data and experience assist the conception and design of future clinical trials. Although inherently difficult, every effort should be made to achieve publication expeditiously.

Recommendations

Design

1. It is essential to specify the hypothesis of the study prior to the start of the trial.
2. Outcome events must be fully defined and the sample size (number of patients to be enrolled or the total duration of enrollment) should be adequate and specified.
3. The methods of statistical testing of the data should be specified in the protocol and should be selected on the basis of the trial design and the hypothesis.
4. A general plan for interim analysis should be included in the protocol; detailed plans must be approved by the DSMC before the first interim analysis.
5. Specific stopping guidelines and procedures should be explicitly stated in the protocol.

Structure

1. The committee structure should be developed to facilitate the conduct of the trial.
2. The DSMC should consist of scientists, including an experienced biostatistician. A clinician expert in the subject of the clinical investigation also should be a member of the committee.
3. No investigator or trial committee member except the DSMC and, rarely, the chairman of the policy or steering committee, should be allowed to see unblinded data until the trial has been completed or there has been a recommendation from the DSMC that the trial should be terminated earlier than planned.
4. The DSMC should be small (three to five members) in order to reduce the possibility of data leakage.
5. Members of the DSMC should not be otherwise involved in the conduct of the trial.
6. Members of the DSMC should be independent of the manufacturer of any intervention tested in the trial.
7. A formal channel for reporting the recommendations of the DSMC must be established.
8. It is essential that major outcome data are collected by an agency independent of the manufacturer of any intervention under test in the trial; it is preferable that the manufacturer not obtain unblinded trial outcome data.

Analysis

1. To preserve the essential balance of the randomization process, the trial results must be analyzed and reported on an intention-to-treat basis. The validity of the intention-to-treat analysis is critical to the integrity of the trial.
2. There is little value and potential for great harm in making a per-protocol or on-treatment analysis. Such analysis should be avoided and, if conducted, should acknowledge the inability to quantify its significance.
3. When the intention-to-treat analysis is inconsistent with good clinical judgment because, for example, too many patients did not stay in their originally assigned limb or compliance to treatment was low, interpretation should be cautious.

4. The trial must be designed from the outset to minimize noncompliance and crossover by careful attention to censorship criteria.

5. Interim analysis should be based on a formal statistical methodology such as group sequential analysis, using an α spending plan.

Interpretation and Presentation

1. When the DSMC recommends early termination, its action is complicated by the incomplete knowledge of recent and unreported event data. The final decision regarding early termination should be resisted until near complete data are at hand.

2. Timely and good quality end point data are essential.

3. Early termination of a trial usually occurs at a point when fluctuation of the trial information reaches a peak or trough; had the trial continued to completion, the statistical outcome might not be appropriately reflected by the point estimate at the time at which the trial was stopped. Dissemination of trial results should include a statement about the possible biases in the point estimate and caution in its use.

4. Extensive extrapolation from data derived from a clinical trial that has been terminated early should be resisted.

5. When a trial is terminated earlier than anticipated, relevant information, especially that relating to a harmful effect, must be speedily disseminated; a letter to a professional journal should not prejudice eventual publication of the full data provided that permission has been obtained from the editor of that journal.

6. Fundamental trial data should be submitted in a timely fashion for publication in an appropriate peer review journal.

Glossary of Terms

Administrative monitoring—Review of interim data related to recruitment, data quality and timeliness, and combined group baseline data for evaluating eligibility and overall event rates to verify design assumptions. Does not include comparison of treatments for benefit or risk.

α Level—The probability of rejecting the hypothesis of no treatment difference when there is no real difference or of a false positive. Also referred to as the probability of a type I error.

α Spending function—A function or rule that governs how the total α level or probability of type I error can be spread out over the interim analysis based on the fraction of information observed. Controls the false-positive error rate with interim analyses to a prespecified level. Can be used to produce sequential monitoring boundaries.

Asymmetric boundary—Critical values plotted over time for evaluating the statistical significance of a statistic comparing treatments. Test statistics larger in absolute values are viewed as statistically significant. Boundary values are not the same for negative (harmful) trends as for positive (beneficial) trends. Less extreme (or smaller in absolute value) critical values are used to detect negative trends for benefit causing the asymmetry with the boundary.

Censoring—The process by which patient outcome data cannot be obtained beyond a specific point in time. For survival data, one only knows that a patient was alive at a specific point in time and one does not know the exact time of death. A censored event is at the time where that patient was lost to follow-up or the study was terminated before the event of interest (eg, death) was observed. An observed death would be an uncensored event.

Controlled clinical trial*—A clinical trial involving one or more test treatments, at least one control treatment, and concurrent enrollment, treatment, and follow-up of all patients in the trial.

Critical Z value—If a standardized test statistic comparing treatments (Z value or normalized statistic) is greater than the critical Z value in absolute value, the treatment difference is declared to be statistically significant. The probability of being more extreme than the critical value by chance (no true treatment difference) is the α level.

Crossover—A patient who does not comply to assigned treatment and begins to adhere to one of the other treatments. Patient may be a drop-in or drop-out, depending on the direction of the crossover.

Data and safety monitoring committee* (DSMC)—Committee of experts whose responsibility is to review periodically accumulating data for evidence of treatment benefit or possible harm.

Data monitoring board (DMB)—Same as DSMC.

Drop-in*—Used to denote a patient in a clinical trial who, although assigned to one study treatment, receives one of the other study treatments in place of or in addition to the assigned treatment.

Drop-out*—A patient enrolled in a clinical trial who is either unwilling or unable to return to the study clinic for regular follow-up visits.

Early termination or stopping*—A condition or provision incorporated into the design of a clinical trial that enables investigators to terminate patient recruitment or treatment if data accumulated during the trial suggest an adverse or beneficial treatment effect.

Efficacy analysis—Analysis comparing two treatments for effect using only subjects who fully complied to assignment treatment and were eligible. To be contrasted with intention-to-treat analysis.

End point*—A primary or secondary event observed in a patient during the course of treatment or follow-up.

Explanatory trial*—Trials that are designed to explain how a treatment works, in which patients are typically analyzed by treatment received and not as assigned.

False-positive error*—The probability of rejecting the null hypothesis of no treatment difference when there is no treatment difference, that is, falsely claiming a treatment difference. Also called type I error.

Group sequential analysis*—A method of interim data analysis that is carried out after enrollment of a specified number of patients or events have been ob-

served to evaluate whether there is evidence of benefit or harm. A special case of sequential analysis.

**Information fraction or time**—The number of events (or outcomes) observed at an interim point in the trial divided by the total expected number of events (or outcomes) at the scheduled termination in the design. A measure of how much of the trial is completed.

**Intention-to-treat analysis**—A method of data analysis in which the primary tabulations and summaries of patient outcome data are by assigned treatment, regardless of compliance to therapy or the protocol.

**Interim data analysis**—Any data analysis done before the trial is finished, for whatever reason, but usually concerned with assessments of treatment effects, either benefit or harm.

**Management trial**—A trial that is designed primarily to provide information on the value of a treatment in normal usage, typically analyzed by intention to treat.

**Manual of operations**—A document or collection of documents that describes the procedures used in a center or set of centers in a clinical trial (eg, manual of operations for study clinics, coordinating center manual of operations).

**Monitoring boundaries**—Critical values for the test statistic plotted over time that evaluates the statistical significance of a statistic comparing treatment effect. Z values larger in absolute values are declared statistical significant.

**Multiple comparisons**—Two or more treatment comparisons, each involving the same outcome measure, are made or are to be made at one designated point in the course of the trial.

**Multiple looks**—Comparisons of the same two treatment groups are made or are to be made at various time points over the course of a trial.

**Multiple outcomes**—A trial involves several different outcome measures, each of which is used or is to be used to make treatment comparisons.

**Nominal P value**—A probability (P) value calculated assuming data is analyzed only once.

**Noncompliance**—Not following a designated procedure or protocol. Usually in reference to some treatment or data collection procedure.

**Null hypothesis**—A hypothesis that postulated no underlying difference in the populations or groups being compared with regard to the factor, trait, characteristic, or condition of interest.

**One-tailed test**—A statistical test of significance based on the null value of no difference versus the set of alternative values that are either to the right or to the left of the null value (eg, the set indicating a positive treatment effect in a clinical trial).

**Outcome variable**—An observation recorded for patients in the trial at one or more time points after enrollment for the purpose of assessing the effects of the study treatments.

**P value**—A value associated with an observed test statistic that indicates the probability that a value as extreme or more extreme that one observed will arise by chance alone in repeated replications of a study.

**Patient compliance**—The degree to which a patient follows a prescribed set of procedures or treatment.

**Phase III trial**—The third and usually final state in testing a new drug in humans. Performed as part of an approved Investigations New Drug Application under Food and Drug Administration guidelines. Usually designed to include random allocation to a control treatment and to experimental treatment.

**Phase IV trial**—Generally, a randomized, controlled trial that is designed to evaluate the long-term safety and efficacy of a drug for a given indication and that is done with Food and Drug Administration approval. Usually carried out after licensure of the drug for that indication.

**Placebo effect**—The effect produced by a placebo, generally measured by comparison of the effect observed in patients receiving the placebo treatment with the effect observed in patients receiving the active treatment.

**Policy and data monitoring board (PDMB)**—An expanded DMB that not only monitors the data but also provides advice on nondata issues in the trial. May also be another name for DMB.

**Post hoc analyses**—Analyses conducted after the results are available that were not defined before the start of the trial. Such analyses are particularly prone to false-positive claims or type I error.

**Power**—The probability of rejecting the null hypothesis when it is false; that is, correctly claiming a treatment difference exists.

**Premature termination**—Early termination of a trial before data are sufficiently strong to be convincing.

**Primary outcome**—The event or condition the trial is designed to ameliorate, delay, or prevent.

**Random**—A chance process in which the occurrence of previous events is of no value in predicting future events. A term used to refer to a sequence of observations, activities, assignments, etc, that is the result of a chance process.

**Randomization**—The process of assigning patient (treatment units) to treatment using a random process, such as by use of a table of random numbers. The process of deriving an order or sequence of items, determinations, specimens, readings, or the like using a random process.

**Run-in phase**—A period before the randomization in a phase III trial to evaluate a subject's ability and willingness to comply with a drug, either placebo or active treatment, or to evaluate if the experimental treatment effects a specified parameter (eg, arrhythmia suppression) in a patient about to be randomized. Goal is to minimize noncompliance and maximize the sensitivity of the randomized study.

**Sample size**—The actual number of patients enrolled in a study or the anticipated number of patients to be enrolled in a study, or the patient recruitment goal.

**Secondary outcome**—An event or condition related to the primary outcome but of less clinical or medical importance than the primary outcome.

**Sequential analysis**—The analysis done after enrollment of a patient, pair of patients, or larger block of patients to determine whether additional patients should be enrolled. The decision is made by observing the test-control difference in observed outcomes. Enrollment of the next patient, pair of patients, or block of patients is carried out if the difference does not exceed prespecified boundary values or critical values.

**Sequential boundary**—Critical values for the test statistic plotted over time, calculated to preserve the α
level or false-positive error rate at a specified level. Can be produced by an $\alpha$ spending function.

**Significance level** — The permissible false-positive type I error level for a test of the null hypothesis with a specified test statistic. The null hypothesis is accepted if the test statistic yield a $P$ value that is larger than the specified level and is rejected if it is equal to or less than this value.

**Steering committee** — A committee responsible for directing the activities of a designated project. Committee responsible for the conduct of the trial and to which all other committees report.

**Stochastic curtailment** — Given trends in the interim data partway through the study, the probability that the null hypothesis will still be rejected or accepted if the trial continued to the planned termination. Used to identify studies where negative trends cannot be reversed with any likelihood to be positive.

**Stopping boundary** — The set of values formed by a line or set of lines (or curves), usually specified before or shortly after the start of patient recruitment, which, if exceeded, indicates the existence of a test-control treatment difference that satisfied certain statistical properties (eg, has $P$ value of less than a certain size). The boundaries will be used as a basis for stopping the trial when developed in conjunction with a sequential design. Same as monitoring boundary.

**Stopping rule** — A rule, usually set before or shortly after the start of patient recruitment, that specified a limit for the observed test-control treatment difference for the primary outcome, which, if exceeded, automatically leads to termination of the test or control treatment, depending on the direction of the observed difference.

**Study protocol** — A narrative document that describes the general design and operating features of a trial. Distinguished from the study manual of operations by its generality and absence of specific details needed for the day-to-day execution of the trial.

**Subgroup analysis** — Any data analysis that focuses on a selected subgroup or patients. Generally, any analysis that is aimed at elucidating treatment differences within a defined subgroup of patients.

**Surrogate outcome** — An outcome based on some laboratory test or measurement that is used instead of a clinical event in the design or analysis of a clinical trial.

**Symmetric boundaries** — Monitoring or sequential boundaries that are equal for evaluating the positive (beneficial) trends and negative (harmful) trends. Critical $Z$ values plotted over time that are equal in absolute value for positive or negative values of the test statistic comparing treatment differences.

**Test of significance** — The evaluation of observed data by calculating a specified test statistic and then deriving the associate $P$ value.

**Test statistics** — The formula or computing algorithm used to carry out a test of significance.

**Two-tailed test** — A statistical test of significance based on the null value of no difference versus the set of all alternative values (ie, those that lie to the right and left of the null value).

**Type I error** — The probability of rejecting the null hypothesis when it is true, usually denoted by the Greek letter $\alpha$. Also referred to as false-positive error.

**Type II error** — The probability of accepting the null hypothesis when it is false, usually denoted by the Greek letter $\beta$. Also referred to as false-negative error.

**Z value** — The standardized test statistic typically obtained by dividing the observed difference by the standard error. A $Z$ value equal to 1.96 in absolute value corresponds to a nominal $P$ value of .05.

**Acknowledgment**

We wish to express our deep gratitude to Pinuccia De Tomasi for typing the several drafts of the manuscript and in coping with the multiple and almost endless revisions made by all the members of the Task Force.

**References**

1. Food and Drug Administration. Final report of the committee for clinical review based on a review of selected medical device applications. March 5, 1993.


Circulation. 1994;89:2892-2907
doi: 10.1161/01.CIR.89.6.2892

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 1994 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/89/6/2892

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Circulation is online at:
http://circ.ahajournals.org/subscriptions/