Chapter 16
Monitoring Committee Structure and Function

The investigator's ethical responsibility to the study participants demands that safety and clinical benefit be monitored during trials. If data partway through the trial indicate that the intervention is harmful to the participants, early termination of the trial should be considered. If these data demonstrate a clear definitive benefit from the intervention, the trial may also be stopped early because continuing would be unethical to the participants in the control group. In addition, if differences in primary and possibly secondary response variables are so unimpressive that the prospect of a clear result is extremely unlikely, it may not be justifiable in terms of time, money, and effort to continue the trial. Also, monitoring of response variables can identify the need to collect additional data to clarify questions of benefit or toxicity that may arise during the trial. Finally, monitoring may reveal logistical problems or issues involving data quality that need to be promptly addressed. Thus, there are ethical, scientific, and economic reasons for interim evaluation of a trial [1–3]. In order to fulfill the monitoring function, the data must be collected and processed in a timely fashion as the trial progresses. Data monitoring would be of limited value if conducted only at a time when all or most of the data had been collected. The specific issues related to monitoring of recruitment, adherence, and quality control are covered in other chapters and will not be discussed here. The monitoring committee process has been described in detail [4] as have case studies representing trials, which were terminated for benefit, harm, or futility [5]. One of the earliest discussions of the basic rationale for data monitoring was included in a report of a committee initiated at the request of the council advisory to the then National Heart Institute and chaired by Bernard Greenberg [3]. This report outlined a clinical trial model depicted in Fig. 16.1, variations of which have been implemented widely by institutes at the National Institutes of Health (NIH). The key components are the Steering Committee, the Statistical and Data Coordinating Center, the Clinics, and the Data Monitoring Committee. Later the pharmaceutical and device industries [6] adopted a modified version of this NIH model, depicted in Fig. 16.2. The main modification was to separate the Statistical Data Coordinating Center into a Statistical Data Analysis Center and a Data Coordinating Center.

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-3-319-18539-2_16

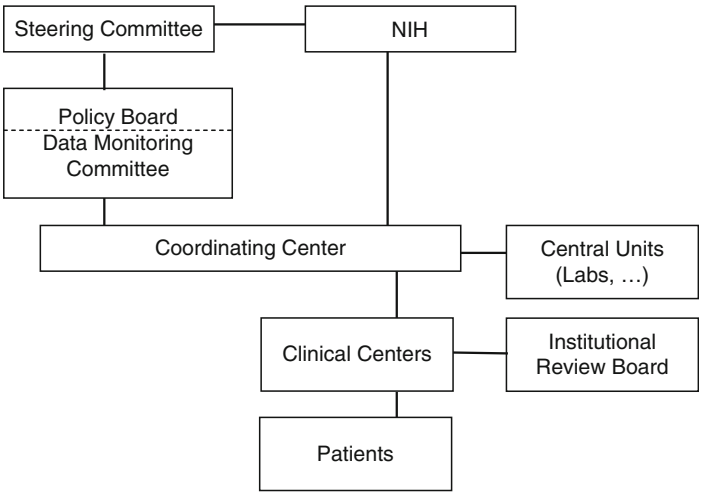

Fig. 16.1 The NIH Clinical Trial Model

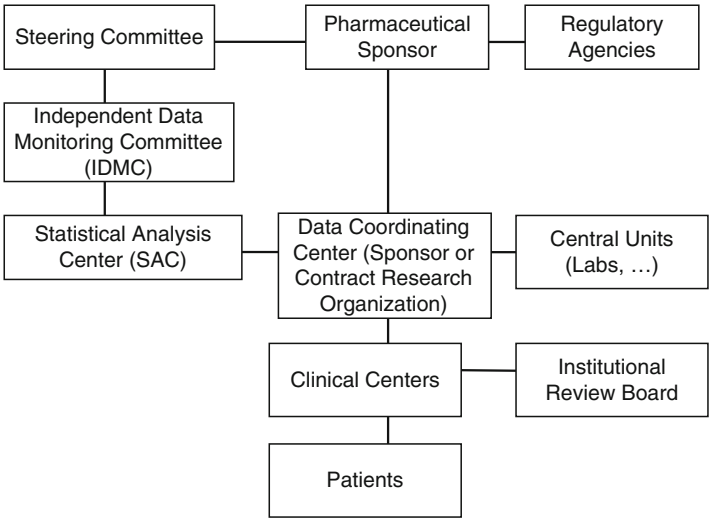

Fig. 16.2 The Industry Modified Clinical Trial Model [6]

Many of the early experiences have been described and formed the basis of current practice [7–34], particularly in trials of cardiovascular disease [35–37]. Over the past decade especially, the number of DMCs has increased dramatically [38]. In 2013, of over 120,000 trials registered in Clintrials.gov, more than 13,000 were interventional trials and 40% of those included a DMC. This suggests over 5,000 DMCs are or have existed in this period of time. The highest DMC utilization was in cardiovascular and oncology trials. Of the 630 trials that were NIH sponsored,

74% had a DMC. For the 55% that were industry sponsored, about a third had a DMC. Some of this difference is reflected in the written policies or guidelines by the NIH and the FDA. The Office of Inspector General (OIG) issued a report in 1998 that reviewed the adequacy of IRB oversight in clinical trials and recommended that the NIH and the FDA provide guidance on when a more focused monitoring committee might be needed. In response, NIH issued a policy statement that was consistent with their longstanding practice in many of their institutes of having an independent DMC for all Phase III randomized trials that they funded [39]. Soon after, the FDA began to develop a guidance document that was issued as a draft in 2001 and finalized in 2006 [40]. The FDA guidance recommended DMCs for trials with high risk patients or high risk or novel interventions, not all Phase III or IV trials conducted by industry.

Prior to the year 2000, the general public was generally not aware of the longstanding practice of data monitoring committees providing oversight to Phase III trials especially. However, the death of a gene transfer patient at a leading research institution changed that [41]. While this patient was not in a Phase III trial, the issues surrounding the case drew attention as to who was responsible for monitoring trials and to whom or what body should such monitoring be reported to. The proximity of this event in time with the NIH and FDA guidance certainly enhanced public awareness and DMC activity came under increased scrutiny by a variety of interested parties. The US Secretary of Health and Human Resources became aware of these events and also reaffirmed the policies and practices of the NIH and the FDA [39, 40, 42]. It has also become clear that for large multicenter trials, an individual IRB, often deluged by sporadic SAEs from the sponsor, is not able to ascertain if there is compelling evidence of risk or benefit based on accumulating data according to (often blinded) treatment in the trial. Thus, that critical role in assuring patient safety can only be played by the DMC.

While all trials need some level of monitoring, many trials such as early phase trials, single center trials, a very simple intervention trial or a trial not involving vulnerable populations, may not need an external monitoring committee. External monitoring, using an independent committee, is used mostly in later phase trials that could lead to change in clinical practice or where special expertise is needed. A survey of monitoring practices conducted by the DAMOCLES group found that the roles of monitoring committees varied widely across trials, sponsors, and regions. While there was a general agreement about the types of trials that needed formal monitoring committees, there was not a uniform practice or policy as to their function [43]. External monitoring committees go by a variety of names such as data and safety monitoring board (DSMB), data and safety monitoring committee (DSMC) or simply Data Monitoring Committee (DMC). In this text, we prefer using DMC since it does not focus on safety when in fact the challenge is to review the risks and the benefits of a new intervention.

The principles and fundamentals expressed in this book reflect the experience of the authors in monitoring numerous trials since the early 1970s.

Fundamental Point

During the trial, response variables need to be monitored for early dramatic benefits or potential harmful effects or futility. Monitoring should be done by a person or group independent of the investigator.

Monitoring Committee

Keeping in mind the scientific, ethical, and economic rationales, data and safety monitoring is not simply a matter of looking at tables or results of statistical analysis of the primary outcome. Rather, it is an active process in which additional tabulations and analysis are suggested and evolve as a result of ongoing review. Monitoring also involves an interaction between the individuals responsible for collating, tabulating, and analyzing the data. For single center studies, the monitoring responsibility could, in principle, be assumed by the investigator. However, he may find himself in a difficult situation. While monitoring the data, he may discover that the results trend in one direction or the other while participants are still being enrolled and/or treated. Presumably, he recruits participants to enter a trial on the basis that he favors neither intervention nor control, a state of clinical equipoise [44]. Knowing that a trend exists may make it difficult for him to continue enrolling participants. It is also difficult for the investigator to follow, evaluate, and care for the participants in an unbiased manner knowing that a trend exists. Furthermore, the credibility of the trial is enhanced if, instead of the investigator, an independent person monitors the response variable data. Because of these considerations, we recommend that for later phase trials the individuals who monitor a clinical trial have no formal involvement with the participants or the investigators, although some disagree [11, 19, 20].

Except for small, short-term studies which could be early or late phase, when one or two knowledgeable individuals may suffice, the responsibility for monitoring response variable data is usually placed with an independent group with expertise in various disciplines [4–6]. The independence protects the members of the monitoring committee from being influenced in the decision-making process by investigators, participants as well as federal or industry sponsors. The committee would usually include experts in the relevant clinical fields or specialties, individuals with experience in the conduct of clinical trials, epidemiologists, biostatisticians knowledgeable in design and analysis, and often for NIH funded trials a bioethicist or participant advocate. While we will describe statistical procedures that are often helpful in evaluating interim results in Chap. 17, the decision process to continue, terminate a trial early, or modify the design is invariably complex and no single statistical procedure is adequate to address all these complexities. Furthermore, no single individual is likely to have all the experiences and expertise to deal with these issues. Thus, as was recommended in the Greenberg Report [3], we suggest that the independent monitoring committee have a multidisciplinary membership.

The first priority of the monitoring committee must be to ensure the safety of the participants in the trial. The second priority is to the investigators and the Institutional Review Boards or ethics committees, who place an enormous trust in the monitoring committee both to protect their participants from harm and to ensure the integrity of the trials. Third, the monitoring committee has a responsibility to the sponsor of the trial, whether it be federal or private. Finally, the monitoring committee provides a service to the drug or device regulatory agency, especially for trials which are utilizing drugs, biologics or devices which still have investigational status.

Although many formats for monitoring committee meetings have been used, one that we recommend allows for exchange of information by all relevant parties and for appropriate confidential and independent review [4, 13]. The format utilizes an open session, a closed session, and an executive session. The open session enables interaction between investigator representatives such as the study principal investigator or chair, the sponsor, the statistical center, the relevant industrial participants, and the monitoring committee. It is uncommon but permissible for a regulatory agency to participate in an open session of the meeting. In this session, issues of participant recruitment, data quality, general adherence, toxicity issues, and any other logistical matter that may affect either the conduct or outcome of the trial are considered in a blinded fashion. After a thorough discussion, the monitoring committee would go into a closed session with DMC members and the statistical reporting statistician or team where analyses of the confidential unblinded outcome data are reviewed. This review would include comparison by intervention groups of baseline variables, primary or secondary variables, safety or adverse outcome variables, adherence measures for the entire group, and examinations of any relevant subgroups. Following this review, the monitoring committee may decide to move into an executive session with DMC members only where decisions about continuation, termination or protocol modification are made. After the DMC review has been completed in closed sessions, they may meet with a representative of the sponsor or investigator leadership to share their recommendations which are usually followed up in a letter. Regardless of how formal, most monitoring committee meetings have such components. One variation is that the DMC meeting begins with a closed session which allows the members to discuss any issues that they want to raise with the investigators and sponsors in the subsequent open session. This discussion may also serve to identify what issues will be central in the second closed session. Thus, the sequence is closed executive, open, closed and ending with an open debriefing session. This particular model, for example, has been used extensively in NIH-sponsored AIDS trials [13].

Before a trial begins and the first monitoring committee meeting is scheduled, it must be decided specifically who attends the various sessions, as outlined above. In general, attendance should be limited to those who are essential for proper monitoring. As noted, it is common for the study principal investigator and sponsor representatives to attend the first open session. If he or she does not provide care for participants in the trial, the principal investigator will sometimes attend the closed session; however, that practice is not recommended. If the study is sponsored by

industry, independence and credibility of the study is best served by no industry attendance at the closed session. Industry sponsored trials that are also managed and analyzed by industry will require a biostatistician from the sponsor who prepares the monitoring report to attend. In such situations the company statistician must have a "firewall" separating her from other colleagues at the company, something that may be difficult to achieve in a way that is convincing to outsiders. However, a common practice for industry-sponsored pivotal Phase III trials is for a separate statistical analysis center to provide the interim analyses and report to the independent monitoring committee [6]. This practice reduces the possibility or perception that interim results are known within the industry sponsor, or the investigator group. Regulatory agency representatives usually do not attend the closed session because being involved in the monitoring decision may affect their regulatory role, should the product be submitted for subsequent approval.

An executive session should involve only the voting members of the monitoring committee, although the independent statistician who provided the data report may also attend. There are many variations of this general outline, including a merger of the closed and executive session since attendance may involve the same individuals.

Most monitoring committees evaluate one, or perhaps two, clinical trials. When a trial is completed, that monitoring committee is dissolved. However, as exemplified by cancer and AIDS, ongoing networks of clinical centers conduct many trials concurrently [11, 13, 18–20, 23]. Cancer trial cooperative groups may conduct trials across several cancer sites, such as breast, colon, lung or head, and neck at any given time, and even multiple trials for a given site depending upon the stage of the cancer or other risk factors. The AIDS trial networks in the United States have likewise conducted trials simultaneously in AIDS patients at different stages of the disease. In these areas, monitoring committees may follow the progress of several trials. In such instances, a very disciplined agenda and a standardized format of the data report enhance the efficiency of the review. If there is a research program of several trials evaluating a new drug, a common DMC may have the advantage of being able to monitor a larger combined experience that will provide for more precise estimates of safety and efficacy. Regardless of the model, the goals and procedures are similar.

Another factor that needs to be resolved before the start of a trial is how the intervention or treatment comparisons will be presented to the monitoring committee. In some trials, the monitoring committee knows the identity of the interventions in each table or figure of the report. In other trials, for two interventions the tables may be labelled as A and B with the identity of A and B remaining blinded until the DMC requests the unblinding on a "need to know" basis. Thus, if there are no trends in either benefit or harm, which is likely to be the case early in a trial, there is no overwhelming reason to know the identity of groups A and B. When trends begin to emerge in either direction, the monitoring committee should have full knowledge of the group identities [45].

In some trials, the monitoring committee is blinded throughout the interim monitoring. In order to achieve this, data reports have complex labeling schemes, such as A versus B for baseline tables, C versus D for primary outcomes, E versus F

for toxicity, and G versus H for various laboratory results. While this degree of blinding may enhance objectivity, it may conflict with the monitoring committee's primary purpose of protecting the participants in the trial from harm or unnecessary continuation. As pointed out by Whitehead [46], the intention of this approach is to deny the DMC a complete picture of the interim data. To assess the progress of the trial, the harm and benefit profile of the intervention must be well understood and the possible tradeoffs weighed. If each group of tables is labeled by a different code, the committee cannot easily assess the overall harm/benefit profile of the intervention, and thus may put participants at unnecessary risk or continue a trial beyond the point at which benefit outweighs risks. Such complex coding schemes also increase the chance for errors in labeling. This practice is not common and not recommended.

No simple formula can be given for how often a monitoring committee should meet. The frequency may vary depending on the phase of the trial [2, 4, 5, 47]. Participant recruitment, follow-up, and closeout phases require different levels of activity. Meetings should not be so frequent that little new data are accumulated in the interim, given the time and expense of convening a committee. If potential toxicity of one of the interventions becomes an issue during the trial, special meetings may be needed. In many long-term clinical trials, the monitoring committees have met regularly at 4- to 6-month intervals, with additional meetings or telephone conferences as needed. In some circumstances, an annual review may be sufficient. However, less frequent review is not recommended since too much time may elapse before a serious adverse effect is uncovered. As described later, another strategy is to schedule monitoring committee meetings when approximately 10, 25, 50, 75, and 100% of the primary outcomes have been observed, or some similar pattern. Thus, there might be an early analysis to check for serious immediate adverse effects with later analyses to evaluate evidence of intervention benefit or harm. Other approaches provide for additional in-between analyses if strong, but as yet non-significant trends emerge. Between committee meetings, the person or persons responsible for collating, tabulating, and analyzing the data assume the responsibility for monitoring unusual situations which may need to be brought to the attention of the monitoring committee.

A monitoring committee often reviews the data for the last time before the data file is closed, and may never see the complete data analysis except as it appears in the publication. There is currently no consistent practice as to whether a monitoring committee meets to review the final complete data set. From one perspective, the trial is over and there is no need for the committee to meet since early termination or protocol modification is no longer an option. From another perspective, the committee has become very familiar with the data, including issues of potential concern, and thus may have insight to share with the investigators and study sponsors. Some trials have scheduled this final meeting to allow the monitoring committee to see the final results before they are presented at a scientific meeting or published.

Based on our experience, we strongly recommend this latter approach. There is a great deal to be gained for the trial and the investigators at a very modest cost. Other remaining issues still need to be resolved. For example, if a worrisome safety trend

or a significant finding is not reported clearly or at all in the primary publication, what are the scientific, ethical, and legal obligations for the monitoring committee to comment on what is not reported? Suppose the committee differs substantially in the interpretation of the primary or safety outcomes? What is the process for resolving differences between it and the investigators or sponsor? These are important questions and the answers are not simple or straightforward, yet are relevant for science and ethics.

Repeated Testing for Significance

In the discussion on sample size (Chap. 8) the issue of testing several hypotheses was raised and referred to as the "multiple testing" problem. Similarly, repeated significance testing of accumulating data is essential to the monitoring function has statistical implications [48–54]. These issues are discussed in more detail in Chap. 17 but the concept of repeated testing is described here. If the null hypothesis, H_0, of no difference between two groups is, in fact, true, and repeated tests of that hypothesis are made at the same level of significance using accumulating data, the probability that, at some time, the test will be called significant by chance alone is larger than the significance level selected in the sample size computation. That is, the rate of incorrectly rejecting the null hypothesis, or making a false positive error, will be larger than what is normally considered acceptable. Trends may emerge and disappear, especially early in the trial, and caution must be used.

In a clinical trial in which the participant response is known relatively soon after entry, the difference in rates between two groups may be compared repeatedly as more participants are added and the trial continues. The usual test statistic for comparing two proportions used is the chi-square test or the equivalent normal test statistic. The null hypothesis is that the true response rates or proportions are equal. If a significance level of 5% is selected and the null hypothesis, H_0, is tested only once, the probability of rejecting H_0 if it is true is 5% by definition. However, if H_0 is tested twice, first when one-half of the data are known and then when all the data are available, the probability of incorrectly rejecting H_0 is increased from 5 to 8% [50]. If the hypothesis is tested five times, with one-fifth of the participants added between tests, the probability of finding a significant result if the usual statistic for the 5% significance level is used becomes 14%. For ten tests, this probability is almost 20%.

In a clinical trial in which long-term survival experience is the primary outcome, repeated tests might be done as more information becomes known about the enrolled participants. Canner [10] performed computer simulations of such a clinical trial in which both the control group and intervention group event rates were assumed to be 30% at the end of the study. He performed 2,000 replications of this simulated experiment. He found that if 20 tests of significance are done within a trial, the chance of crossing the 5% significance level boundaries (i.e., $Z = \pm 1.96$) is, on the average, 35%. Thus, whether one calculates a test statistic for comparing

proportions or for comparing time to event data, repeated testing of accumulating data without taking into account the number of tests increases the overall probability of incorrectly rejecting H_0 and claiming an intervention effect. If the repeated testing continues indefinitely, the null hypothesis is certain to be rejected eventually. Although it is unlikely that a large number of repeated tests will be done, even five or ten can lead to a misinterpretation of the results of a trial if the multiple testing issues are ignored.

A classic illustration of the repeated testing problem is provided by the Coronary Drug Project (CDP) for the clofibrate versus placebo mortality comparison, shown in Fig. 16.3 [10, 54]. This figure presents the standardized mortality comparisons over the follow-up or calendar time of the trial. The two horizontal lines indicate the conventional value of the test statistic, corresponding to a two-sided 0.05 significance level, used to judge statistical significance for studies where the comparison is made just one time. It is evident that the trends in this comparison emerge and weaken throughout, coming close or exceeding the conventional critical values on five monitoring occasions. However, as shown in Fig. 16.4, the mortality curves at the end of the trial are nearly identical, corresponding to the very small standardized statistic at the end of the Fig. 16.3. The monitoring committee for this trial took into consideration the repeated testing problem and did not terminate this trial early just because the conventional values were exceeded.

For ethical, scientific, and economic reasons, all trials must be monitored so as not to expose participants to possible harm unnecessarily, waste precious fiscal and human resources, or miss opportunities to correct flaws in the design [2–5]. However, in the process of evaluating interim results to meet these responsibilities,

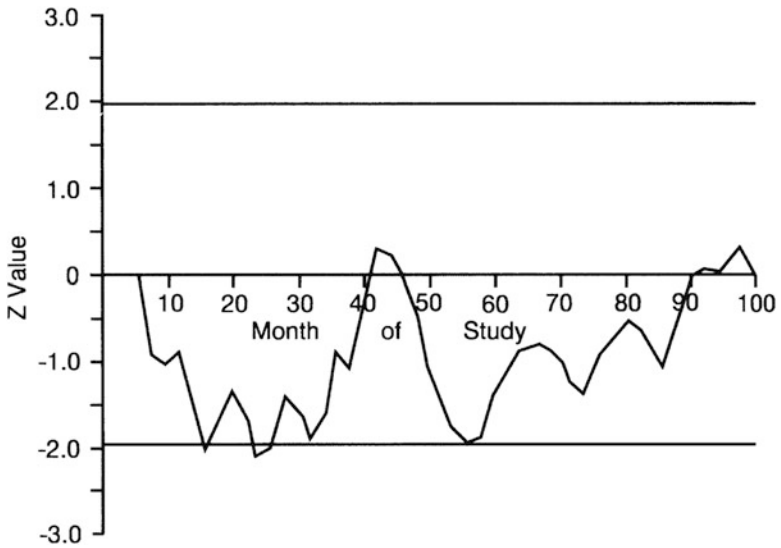

Fig. 16.3 Interim survival analyses comparing mortality in clofibrate- and placebo-treated participants in the Coronary Drug Project. A positive Z value favors placebo [9]

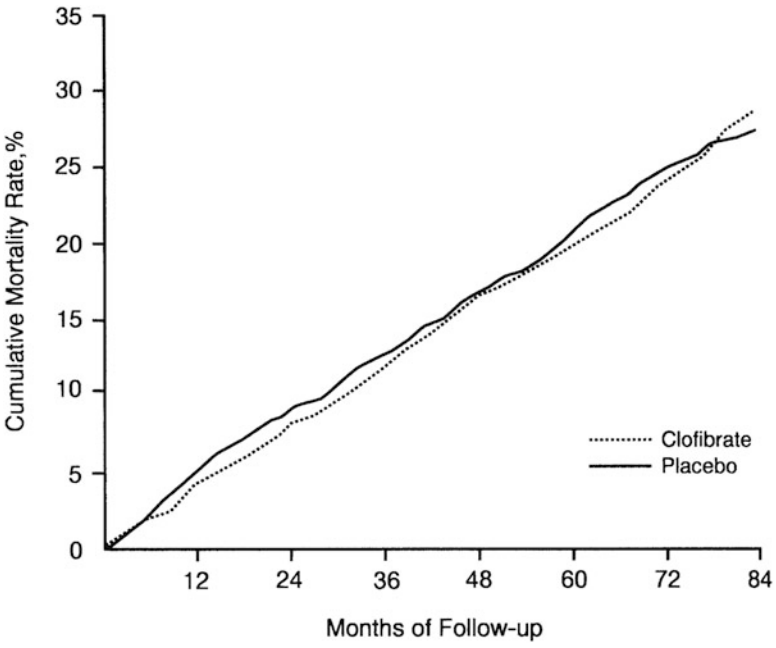

Fig. 16.4 Cumulative mortality curves comparing clofibrate- and placebo treated participants in the Coronary Drug Project [9]

incorrect conclusions can be drawn by overreacting to emerging or non-emerging trends in primary, secondary or adverse effect outcomes. In general, the solution to multiple testing is to adjust the critical value used in each analysis so that the overall significance level for the trial remains at the desired level. It has been suggested that a trial should not be terminated early unless the difference between groups is very significant [2, 4, 5, 55]. More formal monitoring techniques are reviewed in the next chapter, Chap. 17. They include the group sequential methods and stochastic curtailed sampling or conditional power procedures.

Decision for Early Termination

There are five major valid reasons for terminating a trial earlier than scheduled [2, 4, 5, 9, 10]. First, the trial may show serious adverse effects in the entire intervention group or in a dominating subgroup. Second, the trial may indicate greater than expected beneficial effects. Third, it may become clear that a statistically significant difference by the end of the study is improbable, sometimes referred to as being futile. Fourth, logistical or data quality problem may be so severe that correction is not feasible or participant recruitment is far behind and not likely to achieve the target. Fifth, the question posed may have already been

answered elsewhere or may no longer be sufficiently important. A few trials have been terminated because the sponsor decided the trial was no longer a priority but this causes serious ethical dilemmas for investigators and leaves participants having contributed without getting an answer to the posed question.

For a variety of reasons, a decision to terminate a study early must be made with a great deal of caution and in the context of all pertinent data. A number of issues or factors must be considered thoroughly as part of the decision process:

1. Possible differences in prognostic factors between the two groups at baseline.
2. Any chance of bias in the assessment of response variables, especially if the trial is not double-blind.
3. The possible impact of missing data. For example, could the conclusions be reversed if the experience of participants with missing data from one group were different from the experience of participants with missing data from the other group?
4. Differential concomitant intervention and levels of participant adherence.
5. Potential adverse events and outcomes of secondary response variables in addition to the outcome of the primary response variable.
6. Internal consistency. Are the results consistent across subgroups and the various primary and secondary outcome measures? In a multicenter trial, the monitoring committee should assess whether the results are consistent across centers. Before stopping, the committee should make certain that the outcome is not due to unusual experience in only one or two centers.
7. In long-term trials, the experience of the study groups over time. Survival analysis techniques (Chap. 15) partly address this issue.
8. The outcomes of similar trials.
9. The impact of early termination on the credibility of the results and acceptability by the clinical community.

Some trials request the chair of the monitoring committee to review frequently serious adverse events, by intervention, to protect the safety of the participants. While such frequent informal, or even formal, review of the data is also subject to the problems of repeated testing or analyses, the adjustment methods presented are typically not applied. Also, safety may be measured by many response variables. Rather than relying on a single outcome showing a worrisome trend, a profile of safety measures might be required. Thus, the decision to stop a trial for safety reasons can be quite complex.

The early termination of a clinical trial can be difficult [2, 9, 10, 12, 55–60], not only because the issues involved may be complex and the study complicated but also because the final decision often lies with the consensus of a committee. The statistical methods discussed in the next chapter are useful guides in this process but should not be viewed as absolute rules. A compilation of diverse monitoring experiences is available [5]. A few examples are described here to illustrate key points. One of the earlier clinical trials conducted in the United States illustrates how controversial the decision for early termination may be. The University Group Diabetes Program (UGDP) was a placebo-control, randomized, double-blind trial

designed to test the effectiveness of four interventions used in the treatment of diabetes [61–64]. The primary measure of efficacy was the degree of retinal damage. The four interventions were: a fixed dose of insulin, a variable dose of insulin, tolbutamide and phenformin. After the trial was underway, study leaders formed a committee to review accumulating safety data. This committee membership consisted of individuals involved in the UGDP and external consultants. The tolbutamide group was stopped early because the monitoring committee thought the drug could be harmful and did not appear to have any benefit [64]. An excess in cardiovascular mortality was observed in the tolbutamide group as compared to the placebo group (12.7% vs. 4.9%) and the total mortality was in the same direction (14.7% vs. 10.2%). Analysis of the distribution of the baseline factors known to be associated with cardiovascular mortality revealed an imbalance, with participants in the tolbutamide group being at higher risk. This, plus questions about the classification of cause of death, drew considerable criticism. Later, the phenformin group was also stopped because of excess mortality in the control group (15.2% vs. 9.4%) [61]. The controversy led to a further review of the data by an independent group of statisticians. Although they basically concurred with the decisions made by the UGDP monitoring committee [61], the debate over the study and its conclusion continued [63]. This trial certainly highlighted the need for an independent review of the interim data to assess safety.

The decision-making process during the course of the CDP [65] a long-term randomized, double-blind, multicenter study that compared the effect on total mortality of several lipid-lowering drugs (high- and low-dose estrogen, dextrothyroxine, clofibrate, nicotinic acid) against placebo has been reviewed [5, 9, 54, 65, 66]. Three of the interventions were terminated early because of potential adverse effects and no apparent benefit. One of the issues in the discontinuation of the high dose estrogen and dextrothyroxine interventions [65, 67] concerned subgroups of participants. In some subgroups, the interventions appeared to cause increased mortality, in addition to having a number of other adverse effects. In others, the adverse effects were present, but mortality was only slightly reduced or unchanged. The adverse effects were thought to more than outweigh the minimal benefit in selected subgroups. Also, positive subgroup trends in the dextrothyroxine arm were not maintained over time. After considerable debate, both interventions were discontinued. The low dose estrogen intervention [66] was discontinued because concerns over major toxicity. Furthermore, it was extremely improbable that a significant difference in a favorable direction for the primary outcome (mortality) could have been obtained had the study continued to its scheduled termination. Using the data available at the time, the number of future deaths in the control group was projected. This indicated that there had to be almost no further deaths in the intervention group for a significance level of 5% to be reached.

The CDP experience also warns against the dangers of stopping too soon [9, 54]. In the early months of the study, clofibrate appeared to be beneficial, with the significance level reaching or exceeding 5% on five monitoring occasions (Fig. 16.3). However, because of the repeated testing issue described earlier in this chapter, the decision was made to continue the study and closely monitor the results.

The early difference was not maintained, and at the end of the trial the drug showed no benefit over placebo. It is notable that the mortality curves shown in Fig. 16.4 do not suggest the wide swings observed in the interim analyses shown in Fig. 16.3. The fact that participants were entered over a period of time and thus had various lengths of follow-up at any given interim analysis, explains the difference between the two types of analyses. (See Chap. 15 for a discussion of survival analysis.).

Pocock [55] also warns about the dangers of terminating trials too early for benefit, reflecting on a systematic review of trials stopped early [59]. At an early interim analysis, the Candesartan in Heart failure Assessment of Reduction in Mortality and Morbidity (CHARM) trial [68] had a 25% mortality benefit ($p < 0.001$) from candesartan compared to a placebo control, but for a variety of reasons the trial continued and found after a median of 3 years of follow-up only a 9% nonsignificant difference in mortality. Continuing the trial revealed that the early mortality benefit was probably exaggerated and allowed other long-term intervention effects to be discovered. In general, trials stopped early for benefit often do not report in sufficient detail the rationale for early termination and often show implausibly large intervention effects based on only a small number of events [57]. This phenomenon is well recognized [58]. Thus, while there are sound ethical reasons to terminate trials early because of benefit, these decisions must be cautioned by our experience with early trends not being reliable or sustainable. Nevertheless, there is a natural tension between getting the estimate of treatment benefit precise and allowing too many participants to be exposed to the inferior intervention [57]. Statistical methods to be described in the next chapter are useful as guidelines but not adequate as rules and the best approach based on experience is to utilize a properly constituted monitoring committee, charged with weighing the benefits and risks of early termination.

The Nocturnal Oxygen Therapy Trial was a randomized, multicenter clinical trial comparing two levels of oxygen therapy in people with advanced chronic obstructive pulmonary disease [69, 70]. While mortality was not considered as the primary outcome in the design, a strong mortality difference emerged during the trial, notably in one particular subgroup. Before any decision was made, the participating clinical centers were surveyed to ensure that the mortality data were as current as possible. A delay in reporting mortality was discovered and when all the deaths were considered, the trend disappeared. The earlier results were an artifact caused by incomplete mortality data. Although a significant mortality difference ultimately emerged, the results were similar across subgroups in contrast to the results in the earlier review.

Early termination of a subgroup can be especially error prone if not done carefully. Peto and colleagues [71] have illustrated the danger of subgroup analysis by reporting that treatment benefit in ISIS-2 did not apply to individuals born during a certain astrologic sign. Nevertheless, treatment benefits may be observed in subgroups which may be compelling. An AIDS trial conducted by the AIDS Clinical Trial Research Group (ACTG), ACTG-019 [5, 6, 13] indicated that zidovudine (AZT) led to improved outcome in participants who had a low laboratory value (CD4 cell counts under 500, which is a measure of poor immune

response). The results were not significant for participants with a higher CD4 value. Given previous experience with this drug, and given the unfavorable prognosis for untreated AIDS patients, the trial was stopped early for benefit in those with the low CD4 cell count but continued in the rest of the participants.

A scientific and ethical issue was raised in the Diabetic Retinopathy Study, a randomized trial of 1,758 participants with proliferative retinopathy [72, 73]. Each participant had one eye randomized to photocoagulation and the other to standard care. After 2 years of a planned 5 year follow-up, a highly significant difference in the incidence of blindness was observed (16.3% vs. 6.4%) in favor of photocoagulation [74]. Since the long-term efficacy of this new therapy was not known, the early benefit could possibly have been negated by subsequent adverse effects. After much debate, the monitoring committee decided to continue the trial, publish the early results, and allow any untreated eye at high risk of blindness to receive photocoagulation therapy [75]. In the end, the early treatment benefit was sustained over a longer follow-up, despite the fact that some of the eyes randomized to control received photocoagulation. Furthermore, no significant long-term adverse effect was observed.

The Beta-Blocker Heart Attack Trial provided another example of early termination [76, 77]. This randomized placebo control trial enrolled over 3,800 participants with a recent myocardial infarction to evaluate the effectiveness of propranolol in reducing mortality. After an average of a little over 2 years of a planned 3 year follow-up, a mortality difference was observed, as shown in Fig. 16.5. The results were statistically significant, allowing for repeated testing, and would, with high probability, not be reversed during the next year [77]. The data monitoring committee debated whether the additional year of follow-up would add valuable information. It was argued that there would be too few events in the last year of the trial to provide a good estimate of the effect of propranolol treatment in the third and fourth year of therapy. Thus, the committee decided that prompt publication of the observed benefit was more important than waiting for the marginal information yet to be obtained. This trial was one of the early trials to implement group sequential monitoring boundaries discussed in the next chapter and will be used as an example to illustrate the method.

Another example of using sequential monitoring boundaries is found in chronic heart failure trials that evaluated different beta blockers. Common belief had been that administering a beta-blocker drug to a heart failure patient would cause harm, not benefit. Fortunately, early research suggested this belief may have been in error and ultimately four well designed trials were conducted to evaluate the risks and benefits. Three trials were terminated early because of beneficial intervention effect on mortality of 30–35% [78–80]. The fourth trial [81] did not go to completion in part due to the fact that the other three trials had already reported substantial benefits. Details of monitoring in one of the trials, the Metoprolol CR/XL Randomized Trial In Chronic Heart Failure (MERIT-HF) are discussed more fully in Chap. 17.

Some trials of widely used interventions have also been stopped early due to adverse events. One classic example comes from the treatment of arrhythmias following a heart attack. Epidemiological data showed an association between the presence of irregular ventricular heartbeats or arrhythmias and the incidence of

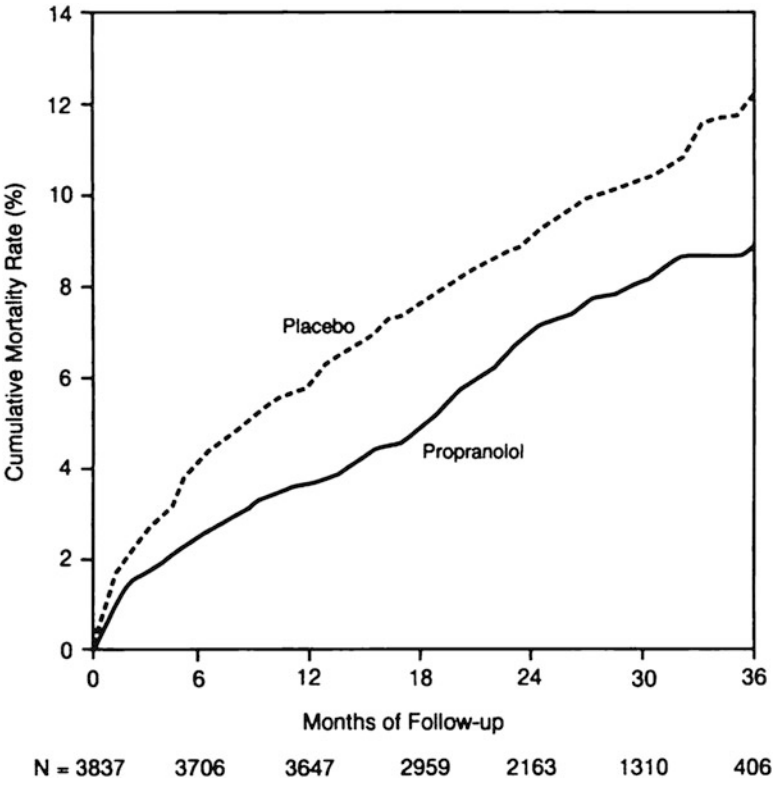

Fig. 16.5 Cumulative mortality curves comparing propranolol and placebo in the Beta-Blocker Heart Attack Trial [75]

sudden death, presumably due to serious arrhythmias. Drugs were developed that suppressed such arrhythmias and they became widely used after approval by the drug regulatory agency for that indication. The Cardiac Arrhythmia Suppression Trial (CAST) was a multicenter randomized double blind placebo-controlled trial evaluating the effects of three such drugs (encainide, flecainide, moricizine) on total mortality and sudden death [82]. Statistical procedures used in CAST to address the repeated testing problem [83, 84] are described in the next chapter. However, the encainide and flecainide arms of the trial were terminated after only 15% of the expected mortality events observed because of an adverse effect (63 deaths in the two active arms vs. 26 deaths in the corresponding placebo arms).

At the first monitoring committee review, the mortality trend in CAST began to appear but the number of events was relatively small [83]. Because the monitoring committee decided no definitive conclusion could be reached on the basis of so few events, it elected to remain blinded to the treatment assignment. However, before the next scheduled meeting, the statistical center alerted the committee that the trends continued and were now nearing the CAST monitoring criteria for stopping.

In a conference call meeting, the monitoring committee became unblinded and learned that the trends were in the unexpected direction, that is, toward harm from the active treatment. A number of confirmatory and exploratory analyses were requested by the monitoring committee and a meeting was held a few weeks later to discuss fully these unexpected results. After a thorough review, the monitoring committee recommended immediate termination of the encainide and flecainide portions of the trial [83]. Results were consistent across outcome variables and participant subgroups, and no biases could be identified which would explain these results. The third arm (moricizine) continued since there were no convincing trends at that time, but it too was eventually stopped due to adverse experiences [85]. The CAST experience points out that monitoring committees must be prepared for the unexpected and that large trends may emerge quickly. Even in this dramatic result, the decision was not simple or straightforward. Many of the issues discussed earlier were covered thoroughly before a decision was reached [83].

Not all negative trends emerge as dramatically as in the CAST. Two other examples are provided by trials in congestive heart failure. Yearly mortality from severe congestive heart failure is approximately 40%. The Prospective RandOmized MIlrinone Survival Evaluation (PROMISE) [36] and the Prospective RandOmized FlosequInan Longevity Evaluation (PROFILE) [35] trials evaluated inotropic agents (milrinone and flosequinone). Both of these drugs had been approved by regulatory agencies for use on the basis of improved exercise tolerance, which might be considered a surrogate response for survival. PROMISE and PROFILE were randomized placebo controlled trials comparing mortality outcomes. Both trials were unexpectedly terminated early due to statistically significant harmful mortality results, even after adjusting for repeated testing of these data. Because severe heart failure has a high mortality rate and the drugs were already in use, it was a difficult decision how long and how much evidence was needed to decide that the intervention was not helpful but was in fact harmful. In both trials, the monitoring committees allowed results to achieve statistical significance since a negative, but nonsignificant trend might have been viewed as evidence consistent with no effect on mortality.

Another trial in acute coronary syndromes, the Thrombin Receptor Antagonist for Clinical Event Reduction in Acute Coronary Syndrome (TRACER) trial, evaluated a thrombin antagonist with a composite outcome of cardiovascular death, myocardial infraction, stroke, recurrent ischemia with rehospitalization or urgent coronary revascularization [86]. The trial of 12,944 patients randomized 1:1 between the thrombin antagonist and placebo was terminated for safety reasons with a positive but non-significant emerging trend in the primary outcome. There were 1,031 primary events in the treated patients and 1,102 in the placebo controls. The secondary composite of cardiovascular death, MI and stroke had 822 vs 910 events ($P = 0.02$). However, the rates of intracranial bleeding was 1.2% vs 0.2% yielding a hazard ratio of 3.39 ($P < 0.001$). The data monitoring group decided that the serious bleeding risks overwhelmed any emerging benefits.

The PROMISE and PROFILE experiences illustrate the most difficult of the monitoring scenarios, the emerging negative trend, but they are not

unique [87–91]. Trials with persistent nonsignificant negative trends may have no real chance of reversing and indicating a benefit from intervention. In some circumstances, that observation may be sufficient to end the trial since if a result falls short of establishing benefit, the intervention would not be used. For example a new expensive or invasive intervention would likely need to be more effective than a standard intervention to be used. In other circumstances, a neutral result may be important, so a small negative trend, still consistent with a neutral result, would argue for continuation. If a treatment is already in clinical use on the basis of other indications, as in the case of the drugs used in PROMISE and PROFILE, an emerging negative trend may not be sufficient evidence to alter clinical practice. If a trial terminates early without resolving convincingly the harmful effects of an intervention, that intervention may still continue to be used. This practice would put future patients at risk, and perhaps even participants in the trial as they return to their usual healthcare system. In that case, the investment of participants, investigators, and sponsors would not have resolved an important question. There is a serious and delicate balance between the responsibility to safeguard the participants in the trial and the responsibility for all concurrent and future patients [87].

Trials may continue to their scheduled termination even though interim results are very positive and persuasive [92] or the intervention and control data are so similar that almost surely no significant results will emerge [93–96]. In one study of antihypertensive therapy, early significant results did not override the need for getting long-term experience with an intensive intervention strategy [92]. Another trial [95] implemented approaches to reduce cigarette smoking, change diet to lower cholesterol, and used antihypertensive medications to lower blood pressure in order to reduce the risk of heart disease. Although early results showed no trends, it was also not clear how long intervention needed to be continued before the applied risk factor modifications would take full effect. It was argued that late favorable results could still emerge. In fact, they did, though not until some years after the trial had ended [96]. In a trial that compared medical and surgical treatment of coronary artery atherosclerosis, the medical care group had such a favorable survival experience that there was little room for improvement by immediate coronary artery bypass graft intervention [94].

The Women's Health Initiative (WHI) was one of the largest and most complex trials ever conducted, certainly in women [97, 98]. This partial factorial trial evaluated three interventions in postmenopausal women: (1) hormone replacement therapy (HRT), (2) a low fat diet, and (3) calcium and vitamin D supplementation. Each intervention, in principle, could affect multiple organ systems, each with multiple outcomes. For example, HRT was being evaluated for its effect on cardiovascular events such as mortality and fatal and non-fatal myocardial infarction. HRT can also affect bone density, the risk of fracture, and breast cancer. The HRT component was also stratified into those with an intact uterus, who received both estrogen and progestin, and those without a uterus who received estrogen alone. The estrogen–progestin arm was terminated early due to increases in deep vein thrombosis, pulmonary embolism, stroke, and breast cancer and a trend toward increased heart disease as shown in Fig. 16.6 although there was a benefit in bone fracture as

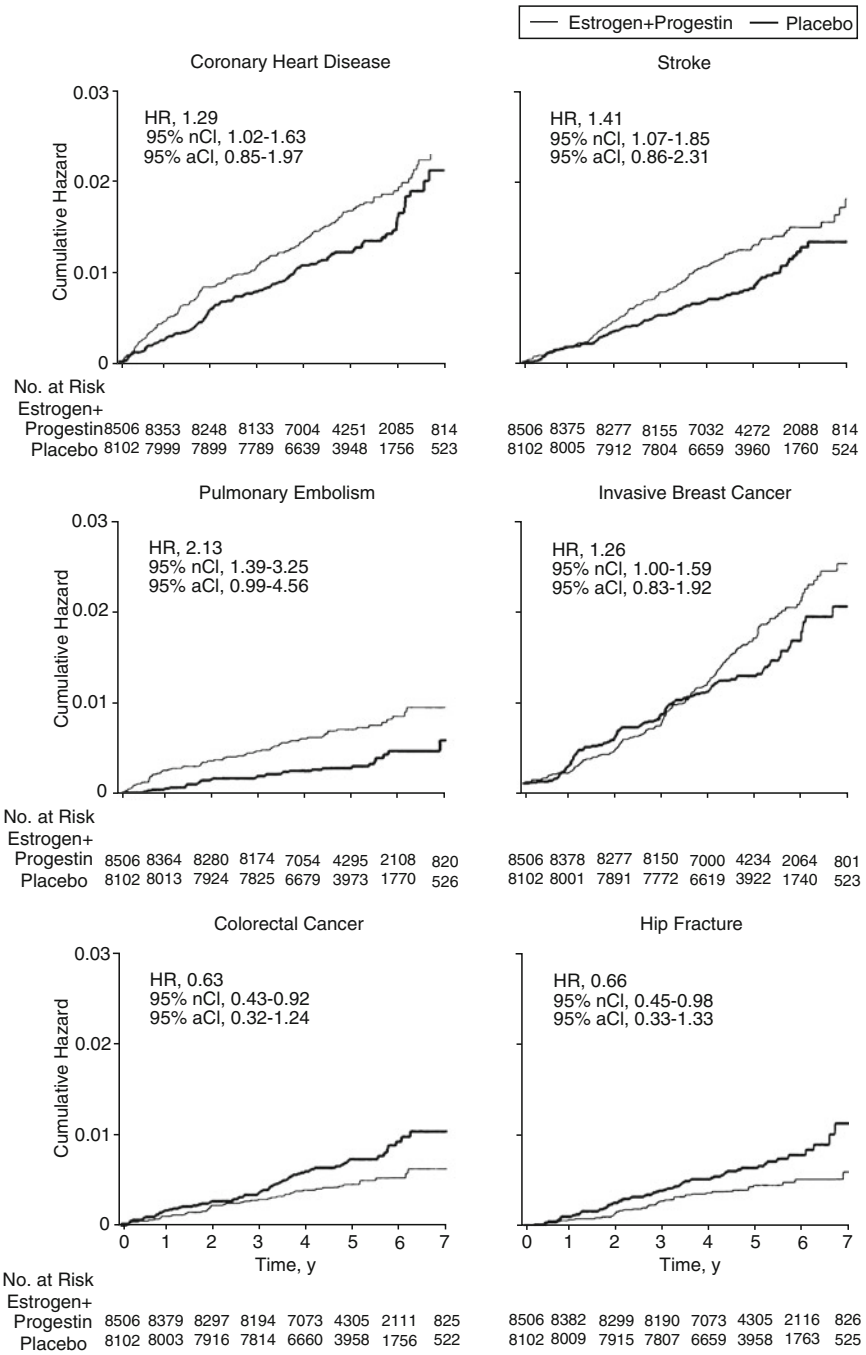

Fig. 16.6 WHI Kaplan-Meier estimates of cumulative hazards for selected clinical outcomes [94]. *HR* hazard ratio, *nCI* nominal confidence interval, *aCI* adjusted confidence interval

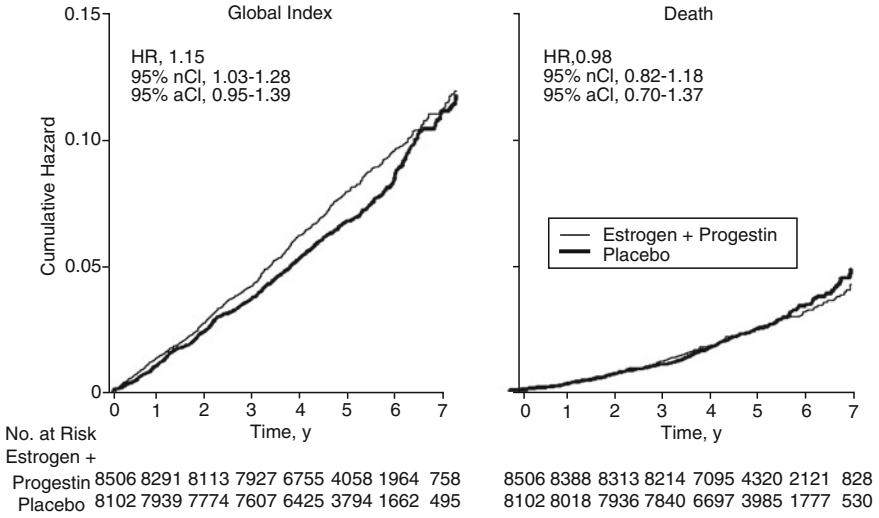

Fig. 16.7 WHI Kaplan-Meier estimates of cumulative hazards for global index and death [94]. *HR* hazard ratio, *nCI* nominal confidence interval, *aCI* adjusted confidence interval

expected [98]. There was no observed difference in total mortality or the overall global index, the composite outcome defined in the protocol, as shown in Fig. 16.7. The WHI is an excellent example of the challenges of monitoring trials with composite outcomes where component trends are not consistent. In such cases, the most important or most clinically relevant component may have to dominate in the decision process, even if not completely specified in the protocol or the monitoring committee charter. Later, the WHI estrogen-alone arm was also terminated, primarily due to increased pulmonary embolus and stroke, though there was no difference in myocardial infarction or total mortality [97]. The formal monitoring process had to account for multiple interventions, multiple outcomes and repeated testing.

A heart failure trial evaluating the drug tezosentan used a stopping criterion that included futility [99]. That is, when there was less than a 10% chance of having a positive beneficial result, the monitoring committee was to alert the investigators and sponsors and recommend termination. In fact, at about two-thirds of the way into the trial, a slightly negative trend was sufficient to make any chance of a beneficial result unlikely and the trial was terminated.

In some instances, a trial may be terminated because the hypothesis being tested has been convincingly answered by other ongoing trials. This was the case with trials evaluating warfarin in the treatment of atrial fibrillation [100]. Between 1985 and 1987, five trials were launched to evaluate warfarin to prevent strokes in participants with atrial fibrillation. Three of the trials were terminated early by 1990, reporting significant reductions in embolic complications. One of the remaining trials was also terminated early, largely due to the ethical aspects of continuing trials when the clinical question being tested has already been answered. The window of opportunity to further evaluate the intervention had closed.

The Justification for the Use of Statin in Prevention: An Intervention Trial Evaluating Rosuvastatin (JUPITER) trial compared a statin agent, which lowers both LDL cholesterol and C-reactive protein, in 17,802 patients with elevated high-sensitivity C-reactive protein levels but without hyperlipidemia [101]. The primary outcome was the occurrence of the combination of myocardial infarction, stroke, arterial revascularization, hospitalization for unstable angina, or death from cardiovascular causes. The trial demonstrated a clear statin effect of lowering LDL even further as well as lowering C-reactive protein levels and demonstrated a corresponding lowering of the primary outcome (hazard ratio (HR) of .56, $p < 0.00001$). Similar reductions were observed for myocardial infarction (HR, 0.46), for stroke (HR, 0.52), for revascularization or unstable angina (HR, 0.53), for the combined end point of myocardial infarction, stroke, or death from cardiovascular causes (HR, 0.53), and for death from any cause (HR, 0.80), all being statistically significant. In addition, all of the major predefined subgroups were consistent. Still, there was criticism that the cardiovascular mortality was not significant even though overall mortality was [102, 103]. This raises the difficult question when using combined outcomes as the primary if each component or at least some components should also be statistically significant before terminating a trial. In general, trials are not designed to demonstrate statistically significant results for any of the components usually due to low events for each of them. To do so would require trials much larger than the one designed. If a component of the combined outcome is of paramount importance, then that outcome should be established as the primary and the trial designed accordingly as described in Chaps. 3 and 8. In the case of the JUPITER trial, the results for the primary outcome and nearly all of its components as well as overall mortality appear to be compelling for a trial to be terminated. This is especially the case when total mortality is significantly reduced in addition to the primary. Another approach to a focus on a component of the primary outcome was in the CHARM program, in which three trials that comprised the overall program each had cardiovascular death and heart failure hospitalization as its primary outcome, and the overall program was powered to assess all-cause mortality. The DMC focused on the effect on mortality in the overall program as the criterion for early termination [68].

As we have already discussed, the decision to terminate a trial is complex. It is never based on a single outcome and may require more than one DMC meeting before a recommendation to terminate is reached. Timing of the recommendation can also be questioned by those external to the trial. In the Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events (ILLUMINATE) trial [104], a new agent torcetrapib, a cholesterylester transfer protein inhibitor that increases HDL cholesterol, was tested to reduce major cardiovascular events. ILLUMINATE was a randomized, double-blind study involving 15,067 patients at high cardiovascular risk, receiving either torcetrapib plus atorvastatin (a statin which lowers LDL cholesterol) or atorvastatin alone. The primary outcome was defined as time to death from coronary heart disease, nonfatal myocardial infarction, stroke, or hospitalization for unstable angina, whichever occurred first. ILLUMINATE clearly demonstrated an increase in HDL, which would be expected to cause a

reduction in cardiovascular risk. However, the trial was terminated early by the DMC and the investigators because of an increased risk of death and cardiac events in patients receiving torcetrapib [104]. To conclude that torcetrapib improved HDL but caused harmful clinical effects was of course disappointing since this was the first testing of an exciting new class of drugs. However, the timing of the recommendation to terminate was challenged by a regulatory agency, which recognized the complexity of such decisions but argued that the trial could and perhaps should have been terminated earlier [105]. Determining at what point there is sufficient and compelling evidence to make a recommendation for termination is often challenging. DMCs do not have the benefit of hindsight while in process of monitoring a trial.

On occasion, trials may have achieved a significant benefit, or show strong trends for benefit, but the DMC recommend early termination for safety reasons. Two trials, the Thrombin Receptor Antagonist in Secondary Prevention of Atherothrombotic Ischemic Events (TRA 2P) trial [106] and TRACER [86] provide examples of such instances. Both trials evaluated a new platelet inhibition agent vorapaxar compared with placebo. TRA 2P had the primary outcome as a composite of death from cardiovascular causes, myocardial infarction, or stroke. TRACER had a composite outcome of death from cardiovascular causes, myocardial infarction, stroke, recurrent ischemia with rehospitalization, or urgent coronary revascularization. Both trials, TRA 2P with 26,449 patients and TRACER with 12,944 patients, had statistically significant beneficial effects in their respective primary outcomes (HR of 0.87 and 0.89). However, the DMCs for both trials recommended early termination and/or modification of the protocol for unacceptable bleeding complications including intracranial hemorrhage.

In all of these studies, the decisions were difficult and involved many analyses, thorough review of the literature, and an understanding of the biological processes. As described above, a number of questions must be answered before serious consideration should be given to early termination. As noted elsewhere, the relationship between clinical trials and practice is very complex and this complexity is evident in the monitoring process [107, 108].

Decision to Extend a Trial

The question of whether to extend a trial beyond its original sample size or planned period of follow-up may arise. Suppose the mortality rate over a 2-year period in the control group is assumed to be 40%. (This estimate may be based on data from another trial involving a similar population.) Also specified is that the sample size should be large enough to detect a 25% reduction due to the intervention, with a two-sided significance level of 5% and a power of 90%. The total sample size is, therefore, approximately 960. However, say that early in the study, the mortality rate in the control group appears somewhat lower than anticipated, closer to 30%. This difference may result from a change in the study population, selection factors in the trial, or new concomitant therapies. If no design changes are made, the

intervention would have to be more effective (30% reduction rather than 25%) for the difference between groups to be detected with the same power. Alternatively, the investigators would have to be satisfied with approximately 75% power of detecting the originally anticipated 25% reduction in mortality. If it is unreasonable to expect a 30% benefit and if a 75% power is unacceptable, the design needs modification. Modifying the design to increase sample size or extend follow-up can inflate type 1 error if it is done with knowledge of the observed intervention effect and not pre-specified. Even when the process is pre-specified, because they may be aware of other data suggesting reasons not to extend the trial, the DMC is usually not involved in such decisions. Changes may be made by a third party either according to a pre-specified plan, or with access to certain summaries of follow-up data, but not to estimates of the intervention effect.

In the above example, given the lower control group mortality rate, approximately 1,450 participants would be required to detect a 25% reduction in mortality, while maintaining a power of 90%. Another option is to extend the length of follow-up, which would increase the total number of events. A combination of these two approaches can also be tried (e.g. [109]). Another approach that has been used [35, 36] is to fix the target of the trial to be a specified number of primary events in the control group or the overall number. This is often referred to as "event driven" trials. If event rates are low, it may take longer follow-up per participant or more randomized participants, or both, to reach the required number of events. In any case, the target is the number of events. In the above situations, only data from the control group or the combined groups are used. No knowledge of what is happening in the intervention group is needed. In our example, if the event rate is closer to 30% than the assumed 40%, then the expected number of events under the null hypothesis, 390, would not be achieved. The trial could achieve the prespecified target number of events by increasing recruitment, increasing the length of follow-up or a combination.

The concept of adaptive designs has already been discussed in Chap. 5. Adaptive designs can be used in trials with overall lower event rates or increased variability, or when emerging trends are smaller than planned for but yet of clinical interest. Modifying the design once the trial is underway due to lower event rates or increased variability is rather straightforward. In a trial of antenatal steroid administration [110], the incidence of infant respiratory distress in the control group was much less than anticipated. Early in the study, the investigators decided to increase the sample size by extending the recruitment phase. In another trial, the protocol specifically called for increasing the sample size if the control group event rate was less than assumed [111]. As described in the sample size chapter, power is the probability of detecting a treatment effect if there truly is an effect. This probability is computed at the beginning of the trial during the design phase. The design goal is to set this probability at a range from 0.80 to 0.95 with an appropriate sample size. Sometimes this probability, or power, is referred to as "unconditional power" to distinguish it from "conditional power" to be described in more detail in the next chapter. Adjustments to sample size based on overall event rates or variability estimates can preserve the power (or unconditional power). No account of emerging trends is used in this recalculation.

The issue of whether the control group event rate or the overall event rate should be used in this sample size reassessment must be considered. It might seem intuitive that the emerging control group event rate should be used since it was the estimated control group rate that was initially used in the sample size calculation, as described in Chap. 8. However, to reveal the control group rate to the investigators may unblind the emerging trend if they are also aware of the overall number of events. The use of the overall event rate would avoid this potential problem. Additionally, there are statistical arguments that under the null hypothesis, the overall rate is the more appropriate one to use because it is likely to be more stable, particularly if the sample size re-estimation is done early in the trial. Many prefer to use the overall event rate, but in either case, this must be decided while the protocol and data monitoring procedures are being developed.

However, modifying the design based on emerging trends is more complicated (see Chap. 5) and will be discussed in more technical detail in the next chapter. Despite the statistical literature for different approaches [112–115] and some criticism [116, 117], only a few applications of this type of adaptive design have been utilized. One such trial is the African-American Heart Failure Trial (A-HeFT) [118], a trial in African Americans with advanced heart failure using a combination of two established drugs. The primary outcome consisted of a weighted score of death, hospitalization, and quality of life. Mortality was among the secondary outcomes. The trial utilized an adaptive design [113] that required the monitoring committee to assess variability of this novel primary outcome and the emerging trend to make sample size adjustment recommendations to the trial leaders. The reason for the adaptive design was that little previous data were available for this combined outcome so estimates of variability were not adequate to compute a reliable sample size. Little experience with the outcome also limited the assessment of potential drug effect on this outcome. A group sequential boundary was established using a Lan–DeMets alpha spending function of the O'Brien–Fleming type (see Chapter 17) for monitoring benefit or harm for the composite outcome, as described in the next chapter. This adaptive procedure was followed as planned and the sample size was increased from 800 to 1,100. Meanwhile, the monitoring committee was observing a mortality trend favoring the combination drug but there was no sequential monitoring plan prespecified for this outcome. The monitoring committee elected to utilize the same sequential boundary specified for the primary composite outcome to monitor mortality. Although not ideal while the trial was ongoing, it was done before the mortality difference became nominally significant. At the last scheduled meeting of the monitoring committee, the difference was nominally significant at the 0.05 level but had not crossed the sequential boundary. The committee decided to conduct an additional review of the data. At that additional review, the mortality difference was nominally significant ($p = 0.01$) and had, in fact, crossed the sequential O'Brien–Fleming boundary. The committee recommended early termination both because of a significant mortality benefit and a primary outcome that was nominally significant, along with a consistency across the components of the composite outcome and relevant subgroups.

While the statistical methods for adaptive designs based on emerging trends to reset the sample size exist, the use of these methods is still evolving. A more technical discussion of specific trend adaptive designs is provided in the next chapter. One concern is whether the application of the pre-specified algorithm, according to the statistical plan, may reveal information about the size and direction of the emerging trend to those blind to the data. These algorithms can be "reverse engineered" to obtain a reasonable estimate of the emerging trend. We know of no example to date where this revelation has caused a problem but in principle this could create bias in participant selection or recruitment efforts or even participant assessment. Thus, mechanisms for implementation of trend adaptive trials are needed that protect the integrity of the trial.

If only the control group event rate and not the intervention effect is used in the recalculation of sample size, then an increase could be recommended when the observed difference between the intervention and control groups is actually larger than originally expected. Thus, in the hypothetical example described above, if early data really did show a 30% benefit from intervention, an increased sample size might not be needed to maintain the desired power of 90%. For this reason, despite the shortcomings of existing adaptive designs, one would not like to make a recommendation about extension without also considering the observed effect of intervention. Computing conditional power is one way of incorporating these results, and some methods in the adaptive design literature have formalized such as approach [112]. Conditional power is the probability that the test statistic will be larger than the critical value, given that a portion of the statistic is already known from the observed data and described in the next chapter. As in other power calculations, the assumed true difference in response variables between groups must be specified. When the early intervention experience is better than expected, the conditional power will be large. When the intervention is doing worse than anticipated, the conditional power will be small. The conditional power concept utilizes knowledge of outcome in both the intervention and control groups and is, therefore, controversial. Nevertheless, the concept attempts to quantify the decision to extend.

Whatever adjustments are made to either sample size or the length of follow-up, they should be made as early in the trial as possible or as part of a planned adaptive design strategy. Early adjustments should diminish the criticism that the trial leadership waited until the last minute to see whether the results would achieve some prespecified significance level before changing the study design. The technical details for the statistical methods to adjust sample size based on interim results are covered in the next chapter.

As mentioned earlier, one challenge in adaptive designs using interim effects that remains unresolved is who should make the calculations and perhaps recommend a sample size increase. The monitoring committee is of course aware of the interim results for the primary outcome but also for the other secondary outcomes and adverse event outcomes as well as overall conduct. The sample size calculations based on the primary events and emerging trends may recommend an increase in sample size but the overall profile of the intervention effects may not support

such an increase. Knowing all of the interim results may place the DMC in an awkward and even ethical dilemma.

Traditionally, DMCs have not been directly involved in the trial design except to possible terminate a trial early. Based on our experience to date, we do not recommend that the monitoring committee engage in the adaptive design if based on emerging trends.

Accelerated Approval Paradigm

One special case of extending a trial is presented in the accelerated approval paradigm as illustrated by the recent FDA guidelines for new interventions for diabetes [119]. Diabetes is a serious disease with fatal and nonfatal consequences. Thus, getting new interventions into clinical practice is a priority. A typical regulatory approval historically may have been based on a drug's ability to lower glucose or HbA1c, which is believed to reduce the risk of diabetes. However, trials in diabetes have raised the issue of whether a new drug might in fact raise cardiovascular (CV) risk. Thus, these new FDA guidelines propose a two-step process. The first step is to rule out a substantial increase in relative risk of 1.8. If and when that criteria is met, the pharmaceutical sponsor can submit their data to get a conditional FDA regulatory approval and be able to market their drug on the condition that they gather additional data to rule out an increase in CV risk of 1.3 or greater. This could be accomplished through two consecutive trials, one smaller trial with approximately 200 CV events to rule out a CV risk of 1.8 followed by a second larger trial with approximately 600 CV events to rule out a CV risk of 1.3. Alternatively, one large trial could be designed such that when the 95% confidence interval excludes a CV risk of 1.8, the monitoring committee could alert the sponsor to that fact and the regulatory submission for drug approval would be initiated. However, the trial would continue to gather for CV data to rule out a risk of 1.3. A problem for trial continuation arises however if the new drug is fact given regulatory conditional approval. At that point, the clinical and private community is now aware of the "interim" results submitted to rule out the CV risk of 1.8, with pressure to share some degree of detail about the results. Whether diabetes patients will continue adhere to their assigned, presumably blinded, trial medication or even agree to be entered if not already randomized is a serious question. Failure to adhere to the assigned therapy will bias the risk assessment towards the null and thus a result favoring a rule out of the 1.3 CV risk.

Employing the two sequential trial approach also has problems. First, it adds to the length of the period where a new drug is under development. Second, knowing the results of the first trial which ruled out a CV risk of 1.8 may influence the profile of the diabetes patients in the second trial, either through physician or patient decision not to participate. For example, more serious risk patients may choose not to be randomized and thus lowering the overall CV risk.

References

1. Baum M, Houghton J, Abrams K. Early stopping rules—clinical perspectives and ethical considerations. *Statist Med* 1994;13:1459–1469.
2. Fleming TR, DeMets DL. Monitoring of clinical trials: issues and recommendations. *Control Clin Trials* 1993;14:183–197.
3. Heart Special Project Committee. Organization, review, and administration of cooperative studies (Greenberg Report): a report from the Heart Special Project Committee to the National Advisory Heart Council, May 1967. *Control Clin Trials* 1988;9:137–148.
4. Ellenberg SS, Fleming TR, DeMets DL. Data Monitoring Committees in Clinical Trials: A Practical Perspective. Wiley, 2003.
5. DeMets DL, Furberg C, Friedman LM. Data monitoring in clinical trials: a case studies approach. New York, NY, Springer, 2006.
6. Fisher MR, Roecker EB, DeMets DL. The role of an independent statistical analysis center in the industry-modified National Institutes of Health model. *Drug Inf J* 2001;35:115–129.
7. Burke G. Discussion of 'early stopping rules—clinical perspectives and ethical considerations'. *Statist Med* 1994;13:1471–1472.
8. Buyse M. Interim analyses, stopping rules and data monitoring in clinical trials in Europe. *Statist Med* 1993;12:509–520.
9. Canner PL. Practical Aspects of Decision-Making In Clinical Trials—The Coronary Drug Project as a Case-Study. *Control Clin Trials* 1981;1:363–376.
10. Canner PL. Monitoring of the data for evidence of adverse or beneficial treatment effects. *Control Clin Trials* 1983;4:467–483.
11. Crowley J, Green S, Liu PY, Wolf M. Data monitoring committees and early stopping guidelines: the Southwest Oncology Group experience. *Statist Med* 1994;13:1391–1399.
12. DeMets DL. Data monitoring and sequential analysis—An academic perspective. *J Acquir Immune Defic Syndr* 1990;3:S124–S133.
13. DeMets DL, Fleming TR, Whitley RJ, et al. The data and safety monitoring board and acquired immune deficiency syndrome (AIDS) clinical trials. *Control Clin Trials* 1995;16:408–421.
14. Ellenberg SS, Myers MW, Blackwelder WC, Hoth DF. The use of external monitoring committees in clinical trials of the National Institute of Allergy and Infectious Diseases. *Statist Med* 1993;12:461–467.
15. Fleming TR, Green SJ, Harrington DP. Considerations for monitoring and evaluating treatment effects in clinical trials. *Control Clin Trials* 1984;5:55–66.
16. Friedman L. The NHLBI model: a 25 year history. *Statist Med* 1993;12:425–431.
17. Geller NL, Stylianou M. Practical issues in data monitoring of clinical trials: summary of responses to a questionnaire at NIH. *Statist Med* 1993;12:543–551.
18. George SL. A survey of monitoring practices in cancer clinical trials. *Statist Med* 1993;12:435–450.
19. Green S, Crowley J. Data monitoring committees for Southwest Oncology Group clinical trials. *Statist Med* 1993;12:451–455.
20. Harrington D, Crowley J, George SL, et al. The case against independent monitoring committees. *Statist Med* 1994;13:1411–1414.
21. Herson J. Data monitoring boards in the pharmaceutical industry. *Statist Med* 1993; 12:555–561.
22. O'Neill RT. Some FDA perspectives on data monitoring in clinical trials in drug development. *Statist Med* 1993;12:601–608
23. Parmar MK, Machin D. Monitoring clinical trials: experience of, and proposals under consideration by, the Cancer Therapy Committee of the British Medical Research Council. *Statist Med* 1993;12:497–504.
24. Pater JL. The use of data monitoring committees in Canadian trial groups. *Statist Med* 1993;12:505–508.

25. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976;34:585.
26. Pocock SJ. Statistical and ethical issues in monitoring clinical trials. *Statist Med* 1993;12:1459–1469.
27. Robinson J. A lay person's perspective on starting and stopping clinical trials. *Statist Med* 1994;13:1473–1477.
28. Rockhold FW, Enas GG. Data monitoring and interim analyses in the pharmaceutical industry: Ethical and logistical considerations. *Statist Med* 1993;12:471–479.
29. Simon R. Some practical aspects of the interim monitoring of clinical trials. *Statist Med* 1994;13:1401–1409.
30. Souhami RL. The clinical importance of early stopping of randomized trials in cancer treatments. *Statist Med* 1994;13:1293–1295.
31. Task Force of the Working Group on Arrythmias of the European Society of Cardiology. The early termination of clinical trials: causes, consequences, and control. With special reference to trials in the field of arrhythmias and sudden death. *Circulation* 1994;89:2892–2907.
32. Walters L. Data monitoring committees: the moral case for maximum feasible independence. *Statist Med* 1993;12:575–580.
33. Williams GW, Davis RL, Getson AJ, et al. Monitoring of clinical trials and interim analyses from a drug sponsor's point of view. *Statist Med* 1993;12:481–492.
34. Wittes J. Behind closed doors: the data monitoring board in randomized clinical trials. *Statist Med* 1993;12:419–424.
35. Packer M, Rouleau J, Swedberg K, et al. Effect of flosequinan on survival in chronic heart failure: preliminary results of the PROFILE study. *Circulation* 1993;88 (Supp I):301.
36. Packer M, Carver JR, Rodeheffer RJ, et al. Effect of Oral Milrinone on Mortality in Severe Chronic Heart Failure. *N Engl J Med* 11-21–1991;325:1468–1475.
37. Packer M, O'Connor CM, Ghali JK, et al. Effect of Amlodipine on Morbidity and Mortality in Severe Chronic Heart Failure. *N Engl J Med* 1996;335:1107–1114.
38. Seltzer J. Clinical Trial Safety-The Goldilocks Dilemma-Balancing Effective and Efficient Safety Monitoring. *Drug Development* 2010;5:8.
39. NIH: NIH policy for data and safety monitoring. NIH Guide.
40. FDA: Guidance for clinical trial sponsors: Establishment and operation of clinical trial data monitoring committees. FDA.
41. Raper SE, Chirmule N, Lee FS, et al. Fatal systemic inflammatory response syndrome in a ornithine transcarbamylase deficient patient following adenoviral gene transfer. *Mol Genet Metab* 2003;80:148–158.
42. Shalala D: Protecting research subjects-what must be done. *N Engl J Med* 2000;343:808–810.
43. Clemens F, Elbourne D, Darbyshire J, Pocock S. Data monitoring in randomized controlled trials: surveys of recent practice and policies. *Clin Trials* 2005;2:22–33.
44. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987.
45. Meinert CL. Masked monitoring in clinical trials--blind stupidity? *N Engl J Med* 1998;338:1381–1382.
46. Whitehead J. On being the statistician on a Data and Safety Monitoring Board. *Statist Med* 12-30–1999;18:3425–3434.
47. Li ZQ, Geller NL. On the Choice of Times for Data Analysis in Group Sequential Clinical Trials. *Biometrics* 1991;47:745–750.
48. Anscombe FJ. Sequential medical trials. *J Am Stat Assoc* 1963;58:365–383.
49. Armitage P. Restricted sequential procedures. *Biometrika* 1957;9–26.
50. Armitage P, McPherson CK, Rowe BC. Repeated Significance Tests on Accumulating Data. *J R Stat Soc Ser A* 1969;132:235–244.
51. Bross I. Sequential medical plans. *Biometrics* 1952;8:188–205.
52. Robbins H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 1952;58:527–535.

53. Robbins H. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* 1970;1397–1409.
54. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
55. Pocock SJ. When to stop a clinical trial. *BMJ* 1992;305:235.
56. DeMets DL. Stopping Guidelines Vs Stopping Rules—A Practitioners Point of View. *Commun Stat Theory Methods* 1984;13:2395–2417.
57. Freidlin B, Korn EL. Stopping clinical trials early for benefit: impact on estimation. *Clin Trials* 2009;6:119–125.
58. Goodman SN. Stopping trials for efficacy: an almost unbiased view. *Clin Trials* 2009;6:133–135.
59. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. *JAMA* 2005;294:2203.
60. Pocock SJ. When (not) to stop a clinical trial for benefit. *JAMA* 2005;294:2228–2230.
61. Report of the committee for the assessment of biometric aspects of controlled trials of hypoglycemic agents. *JAMA* 1975;231:583–608.
62. Knatterud GL, Meinert CL, Klimt CR, et al. Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: IV. A preliminary report on phenformin results. *JAMA* 1971;217:777–784.
63. Kolata GB. Controversy over study of diabetes drugs continues for nearly a decade. Science (New York, NY) 1979;203:986.
64. Meinert CL, Knatterud GL, Prout TE, Klimt CR. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. II. Mortality results *Diabetes* 1970;19:789–830.
65. The Coronary Drug Project Research Group. The coronary drug project: Initial findings leading to modifications of its research protocol. *JAMA* 1970;214:1303–1313.
66. The Coronary Drug Project Research Group. The coronary drug project: Findings leading to discontinuation of the 2.5-mg/day estrogen group. *JAMA* 1973;226:652–657.
67. The Coronary Drug Project Research Group. The coronary drug project: Findings leading to further modifications of its protocol with respect to dextrothyroxine. *JAMA* 1972;220:996–1008.
68. Pocock SJ, Wang D, Wilhelmsen L, Hennekens CH. The data monitoring experience in the Candesartan in Heart Failure Assessment of Reduction in Mortality and morbidity (CHARM) program. *Am Heart J* 2005;149:939–943.
69. DeMets DL, Williams GW, Brown Jr BW. A case report of data monitoring experience: The Nocturnal Oxygen Therapy Trial. *Control Clin Trials* 1982;3:113–124.
70. Nocturnal Oxygen Therapy Trial Group. Continuous or Nocturnal Oxygen Therapy in Hypoxemic Chronic Obstructive Lung DiseaseA Clinical Trial. *Ann Intern Med* 1980;93:391–398.
71. ISIS-2 Collaborative Group. Randomised Trial Of Intravenous Streptokinase, Oral Aspirin, Both, Or Neither Among 17 187 Cases Of Suspected Acute Myocardial Infarction: ISIS-2. *Lancet* 1988;332:349–360.
72. Diabetic Retinopathy Study Research Group: Preliminary report on effects of photocoagulation therapy. *Am J Ophthalmol* 1976;81:383–396.
73. Diabetic Retinopathy Study Research Group. Diabetic retinopathy study. Report Number 6. Design, methods, and baseline results. *Invest Ophthalmol Vis Sci* 1981;21:1–226.
74. Diabetic Retinopathy Study Research Group. Photocoagulation treatment of proliferative diabetic retinopathy: the second report of diabetic retinopathy study findings. *Ophthalmology* 1978;85:82–106.
75. Ederer F, Podgor MJ. Assessing possible late treatment effectsin stopping a clinical trial early: A case study. Diabetic retinopathy study report no. 9. *Control Clin Trials* 1984;5:373–381.

76. Beta-Blocker Heart Attack Trial Research Group. A randomized trial of propranolol in patients with acute myocardial infarction: I. mortality results. *JAMA* 1982;247:1707–1714.
77. DeMets DL, Hardy R, Friedman LM, Gordon Lan KK. Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Control Clin Trials* 1984;5:362–372.
78. CIBIS-II Investigators and Committees. The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial. *Lancet* 1999;353:9–13.
79. MERIT HF Study Group. Effect of Metoprolol CR/XL in chronic heart failure: Metoprolol CR/XL Randomized Interventional Trial in congestive heart failure.(MERIT-HF). *Lancet* 1999;353:2001–2007.
80. Packer M, Coats AJS, Fowler MB, et al. Effect of Carvedilol on Survival in Severe Chronic Heart Failure. *N Engl J Med* 2001;344:1651–1658.
81. Beta-Blocker Evaluation of Survival Trial Investigators. A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *N Engl J Med* 2001;344:1659.
82. 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* 1989;321:406–412.
83. Friedman LM, Bristow JD, Hallstrom A, et al. Data monitoring in the cardiac arrhythmia suppression trial. *Online J Curr Clin Trials* 1993;79.
84. Pawitan Y, Hallstrom A. Statistical interim monitoring of the cardiac arrhythmia suppression trial. *Statist Med* 1990;9:1081–1090.
85. Cardiac Arrhythmia Suppression Trial Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992;327:227–233.
86. Tricoci P, Huang Z, Held C, et al. Thrombin-Receptor Antagonist Vorapaxar in Acute Coronary Syndromes. *N Engl J Med* 2011;366:20–33.
87. DeMets DL, Pocock SJ, Julian DG. The agonising negative trend in monitoring of clinical trials. *Lancet* 1999;354:1983–1988.
88. Furberg CD, Campbell R, Pitt B. ACE Inhibitors after Myocardial Infarction. *N Engl J Med* 1993;328:966–969.
89. Pater JL. Timing the collaborative analysis of three trials comparing 5-FU plus folinic acid (FUFA) to surgery alone in the management of resected colorectal cancer: A National Cancer Institute of Canada Clinical trials group (NCIC-CTG) perspective. *Statist Med* 1994;13:1337–1340.
90. Swedberg K, Held P, Kjekshus J, et al. Effects of the early administration of enalapril on mortality in patients with acute myocardial infarction: results of the Cooperative New Scandinavian Enalapril Survival Study II (CONSENSUS II). *N Engl J Med* 1992;327:678–684.
91. Sylvester R, Bartelink H, Rubens R. A reversal of fortune: practical problems in the monitoring and interpretation of an EORTC breast cancer trial. *Statist Med* 1994;13:1329–1335.
92. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the hypertension detection and follow-up program: I. Reduction in mortality of persons with high blood pressure, including mild hypertension. *JAMA* 1979;242:2562–2571.
93. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
94. CASS Principle Investigators and Their Associates. Coronary artery surgery study (CASS): a randomized trial of coronary artery bypass surgery. Survival data. *Circulation* 1983;68:939–950.
95. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial: Risk factor changes and mortality results. *JAMA* 1982;248:1465–1477.
96. Multiple Risk Factor Intervention Trial Research Group. Mortality after 16 years for participants randomized to the Multiple Risk Factor Intervention Trial. *Circulation* 1996;94:946–951.

97. The Women's Health Initiative Steering Committee. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: the Women's Health Initiative randomized controlled trial. *JAMA* 2004;291:1701–1712.
98. Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the women's health initiative randomized controlled trial. *JAMA* 2002;288:321–333.
99. McMurray JJ, Teerlink JR, Cotter G, et al. Effects of tezosentan on symptoms and clinical outcomes in patients with acute heart failure: the VERITAS randomized controlled trials. *JAMA* 2007;298:2009–2019.
100. Tegeler CH, Furberg CD. Lessons from warfarin trials in atrial fibrillation: Missing the window of opportunity; in Data Monitoring in Clinical Trials: Springer, 2006, pp 312–319.
101. Ridker PM, Danielson E, Fonseca FA, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med* 2008;359:2195.
102. Ridker PM. The JUPITER trial results, controversies, and implications for prevention. *Circ Cardiovasc Qual Outcomes* 2009;2:279–285.
103. Voss E, Rose CP, Biron P. JUPITER, a statin trial without cardiovascular mortality benefit. *Circ Cardiovasc Qual Outcomes* 2009;2:279–285.
104. Barter PJ, Caulfield M, Eriksson M, et al. Effects of Torcetrapib in Patients at High Risk for Coronary Events. *N Engl J Med* 2007;357:2109–2122.
105. Hedenmalm K, Melander H, Alvan G. The conscientious judgement of a DSMB—statistical stopping rules re-examined. *Eur J Clin Pharmacol* 2008;64:69–72.
106. Morrow DA, Braunwald E, Bonaca MP, et al. Vorapaxar in the Secondary Prevention of Atherothrombotic Events. *N Engl J Med* 2012;366:1404–1413.
107. Liberati A. Conclusions. 1: The relationship between clinical trials and clinical practice: The risks of underestimating its complexity. *Statist Med* 1994;13:1485–1491.
108. O'Neill RT. Conclusions. 2: The relationship between clinical trials and clinical practice: The risks of underestimating its complexity. *Statist Med* 1994;13:1493–1499.
109. Brancati FL, Evans M, Furberg CD, et al. Midcourse correction to a clinical trial when the event rate is underestimated: the Look AHEAD (Action for Health in Diabetes) Study. *Clin Trials* 2012;9:113–124.
110. Collaborative Group on Antenatal Steroid Therapy. Effect of antenatal dexamethasone administration on the prevention of respiratory distress syndrome. *Am J Obstet Gynecol* 1981;141:276–287.
111. The MIAMI Trial Research Group. Metoprolol in acute myocardial infarction (MIAMI). A randomised placebo-controlled international trial. *Eur Heart J* 1985;6:199–226.
112. Chen YH, DeMets DL, Lan KK. Increasing the sample size when the unblinded interim result is promising. *Stat Med* 2004;23:1023–1038.
113. Cui L, Hung HMJ, Wang SJ. Modification of Sample Size in Group Sequential Clinical Trials. *Biometrics* 1999;55:853–857.
114. Lan KKG, Trost DC. Estimation of parameters and sample size re-estimation. Proceedings—Biopharmaceutical Section American Statistical Association, 48-51. 1997. American Statistical Association.
115. Proschan MA, Liu Q, Hunsberger S. Practical midcourse sample size modification in clinical trials. *Control Clin Trials* 2003;24:4–15.
116. Fleming TR. Standard versus adaptive monitoring procedures: a commentary. *Statist Med* 2006;25:3305–3312.
117. Tsiatis AA, Mehta C. On the Inefficiency of the Adaptive Design for Monitoring Clinical Trials. *Biometrika* 2003;90:367–378.
118. Taylor AL, Ziesche S, Yancy C, et al. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med* 2004;351:2049–2057.
119. FDA: Guidance for Industry Diabetes Mellitus—Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes. FDA.