Chapter 18
Issues in Data Analysis

The analysis of data obtained from a clinical trial represents the outcome of the planning and implementation already described. Primary and secondary questions addressed by the clinical trial can be tested and new hypotheses generated. Data analysis is sometimes viewed as simple and straightforward, requiring little time, effort, or expense. However, careful analysis usually requires a major investment in all three. It must be done with as much care and concern as any of the design or data-gathering aspects. Furthermore, inappropriate statistical analyses can introduce bias, result in misleading conclusions and impair the credibility of the trial.

In the context of clinical trials, the term bias has several senses. One is what we might call experimenter bias. This sense applies to a difference in the behavior, conscious or unconscious, of investigators depending on what they believe about the intervention. For example, in an unblinded trial, or a trial in which the intervention assignment can be guessed, the investigator may treat a participant or data from a participant differently depending on whether she believes that participant has received the experimental intervention. These differences in behavior can lead to a second sense of the term bias, one with a more technical definition which we may call estimation bias. If the goal of the trial is to estimate how the intervention affects an outcome measured in a specified population, bias is any difference between that estimate and the true effect which is not attributable to random variation.

Even in a randomized, double-blind clinical trial which has been conducted without any experimenter bias, the estimated effect can be biased by excluding randomized participants or observed outcomes, by inappropriate choice of analytic techniques, and by missing or poor quality data caused by mechanisms which were not the same in the intervention and control groups. This chapter will focus on how to avoid introducing estimation bias into the analysis of clinical trial data.

Several introductory textbooks of statistics [1–8] provide excellent descriptions for many basic methods of analysis. Chapter 15 presents essentials for analysis of survival data, since they are frequently of interest in clinical trials. This chapter will focus on some issues in the analysis of data which seem to cause confusion in the

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-3-319-18539-2_18

medical research community. Some of the proposed solutions are straightforward; others require judgment. They reflect a point of view toward bias developed by the authors of this text and many colleagues in numerous collaborative efforts over three to four decades. Whereas some [9–12] have taken similar positions, others [13–16] have opposing views on several issues, and some published trials are more consistent with these opposing views (e.g. [17]).

The analytic approaches discussed here primarily apply to late phase (III and IV) trials. Various exploratory analysis approaches may be entirely reasonable in early phase (I and II) studies where the goal is to obtain information and insight to design better subsequent trials. However, some the fundamentals presented may still be of value in these early phase trials. We have used early examples that were instrumental in establishing many of the analytic principles and added new examples which reinforce them. However, given the multitude of clinical trials, it is not possible to include even a small proportion of the many excellent examples that exist.

Fundamental Point

Removing randomized participants or observed outcomes from analysis and subgrouping on the basis of outcome or other response variables can lead to biased results. Those biases can be of unknown magnitude and direction.

Which Participants Should Be Analyzed?

In the context of clinical trials the term ‘withdrawal’ is used in different ways. For this chapter, ‘withdrawing a participant’ generally means removing from an analysis the data contributed by a participant who has been randomized and perhaps followed for some length of time. A common and related meaning of the term ‘withdrawal’ refers to a participant who is randomized but from whom follow-up data is deliberately not collected, or only partially collected, as a result of decisions made by the study investigators. Yet another meaning indicates someone who is randomized but refuses to continue participating in the trial. The term ‘excluded’ can also be ambiguous, sometimes referring to a participant who does not meet eligibility criteria during screening, sometimes to a participant’s data which was not used in an analysis. Since removing or not using data collected from a randomized participant can lead to bias, the question of which participants’ data should be analyzed is an important one. This chapter has adopted, in part, the terminology used by Peto and colleagues [12] to classify participants according to the nature and extent of their participation.

Discussion about which participants are to be included in the data analysis often arises in clinical trials. Although a laboratory study may have carefully regulated

experimental conditions, even the best designed and managed clinical trial cannot be perfectly implemented. Response variable data may be missing, adherence to protocol may not be complete, and some participants, in retrospect, will not have met the entrance criteria. Some investigators may, after a trial has been completed, be inclined to remove from the analysis participants who did not meet the eligibility criteria or did not follow the protocol perfectly. In contrast, others believe that once a participant is randomized, that participant should always be followed and included in the analysis.

The *intention-to-treat principle* states that all participants randomized and all events, as defined in the protocol, should be accounted for in the primary analysis [12]. This requirement is stated in the International Conference on Harmonisation and FDA guidelines [18, 19]. There are often proposed "modified intention-to-treat" analyses, or "per protocol" or "on treatment" analyses, that suggest that only participants who received at least some of the intervention should be included. However, as we will discuss, any deviations from strict intention-to-treat offer the potential for bias and should be avoided, or at a minimum presented along with an intention-to-treat analysis. Many published analyses claim to have followed the intention-to-treat principle yet do not include all randomized participants and all events. Although the phrase is widely used, "per protocol" analysis suggests that the analysis is the one preferred in the trial's protocol. For such analyses we think that "on treatment" analysis more accurately reflects what is done.

Exclusions refer to people who are screened as potential participants for a randomized trial but who do not meet all of the entry criteria and, therefore, are not randomized. Reasons for exclusion might be related to age, severity of disease, refusal to participate, or any of numerous other determinants evaluated before randomization. Since these potential participants are not randomized, their exclusion does not bias any intervention-control group comparison (sometimes called *internal validity*). Exclusions do, however, influence the broader interpretation and applicability of the results of the clinical trial (*external validity*). In some circumstances, follow-up of excluded people, as was done in the Women's Health Initiative [20, 21], can be helpful in determining to what extent the results can be generalized. If the event rate in the control group is considerably lower than anticipated, an investigator may want to determine whether most high risk people were excluded or whether she was incorrect in her initial assumption.

Withdrawals from analysis refer to participants who have been randomized but are deliberately excluded from the analysis. As the fundamental point states, omitting participants from analyses can bias the results of the study [22]. If participants are withdrawn, the burden rests with the investigator to convince the scientific community that the analysis has not been biased. However, this is essentially impossible, because no one can be sure that participants were not differentially withdrawn from the study groups. Differential withdrawal can occur even if the number of omitted participants is the same in each group, since the reasons for withdrawal in each group may be different and consequently their risk of primary, secondary and adverse events may be different. As a result, the participants

remaining in the randomized groups may not be comparable, undermining one of the reasons for randomization.

Many reasons are given for not including certain participants' data in the analysis, among them ineligibility and nonadherence.

Ineligibility

A previously common cited reason for withdrawal is that some participants did not meet the entry criteria, a protocol violation unknown at the time of enrollment. Admitting unqualified participants may be the result of a simple clerical error, a laboratory error, a misinterpretation, or a misclassification. Clerical mistakes such as listing wrong sex or age may be obvious. Other errors can arise from differing interpretation of diagnostic studies such as electrocardiograms, x-rays, or biopsies. It is not difficult to find examples in earlier literature [23–31]. The practice of withdrawal for ineligibility used to be common, but appears to be less frequent now, at least in papers published in major journals.

Withdrawals for ineligibility can involve a relatively large number of participants. In an early trial by the Canadian Cooperative Study Group [30], 64 of the 649 enrolled participants (10%) with stroke were later found to have been ineligible. In this four-armed study, the numbers of ineligible participants in the study groups ranged from 10 to 25. The reasons for the ineligibility of these 64 participants were not reported, nor were their outcome experiences. Before cancer cooperative groups implemented phone-in or electronic eligibility checks, 10–20% of participants entered into a trial may have been ineligible after further review. By taking more careful care at the time of randomization, the number of ineligible participants was reduced to a very small percent [32]. Currently, web based systems or Interactive Voice Recording Systems are used for multicenter and multinational clinical trials [33]. These interactive systems can lead clinic staff through a review of key eligibility criteria before randomization is assigned, cutting down on the ineligibility rate. For example, several trials employed these methods [34–38].

A study design may require enrollment within a defined time period following a qualifying event. Because of this time constraint, data concerning a participant's eligibility might not be available or confirmable at the time the decision must be made to enroll him. For example, the Beta-Blocker Heart Attack Trial looked at a 2–4 year follow-up mortality in people administered a beta-blocking drug during hospitalization for an acute myocardial infarction [23]. Because of known variability in interpretation, the protocol required that the diagnostic electrocardiograms be read by a central unit. However, this verification took several weeks to accomplish. Local investigators, therefore, interpreted the electrocardiograms and decided whether the patient met the necessary criteria for inclusion. Almost 10% of the enrolled participants did not have their myocardial infarction confirmed according to a central reading, and were "incorrectly" randomized. The question then arose: Should the participants be kept in the trial and included in the analysis of the

response variable data? The Beta-Blocker Heart Attack Trial protocol required follow-up and analysis of all randomized participants. In this case, the observed benefits from the intervention were similar in those eligible as well as in those "ineligible."

A more complicated situation occurs when the data needed for enrollment cannot be obtained until hours or days have passed, yet the study design requires initiation of intervention before then. For instance, in the Multicenter Investigation for the Limitation of Infarct Size [26], propranolol, hyaluronidase, or placebo was administered shortly after participants were admitted to the hospital with possible acute myocardial infarctions. In some, the diagnosis of myocardial infarction was not confirmed until after electrocardiographic and serum enzyme changes had been monitored for several days. Such participants were, therefore, randomized on the basis of a preliminary diagnosis of infarction. Subsequent testing may not have supported the initial diagnosis. Another example of this problem involves a study of pregnant women who were likely to deliver prematurely and, therefore, would have children who were at a higher than usual risk of being born with respiratory distress syndrome [24]. Corticosteroids administered to the mother prior to delivery were hypothesized to protect the premature child from developing this syndrome. Although, at the time of the mother's randomization to either intervention or control groups, the investigator could not be sure that the delivery would be premature, she needed to make a decision whether to enroll the mother into the study. Other examples include trials where thrombolytic agents are being evaluated in reducing mortality and morbidity during and after an acute myocardial infarction. In these trials, agents must be given rapidly before diagnosis can be confirmed [39].

To complicate matters still further, the intervention given to a participant can affect or change the entry diagnosis. For example, in the above mentioned study to limit infarct size, some participants without a myocardial infarction were randomized because of the need to begin intervention before the diagnosis was confirmed. Moreover, if the interventions succeeded in limiting infarct size, they could have affected the electrocardiogram and serum enzyme levels. Participants in the intervention groups with a small myocardial infarction may have had the infarct size reduced or limited and therefore appeared not to have had a qualifying an infarction. Thus, they would not seem to have met the entry criteria. However, this situation could not exist in the placebo control group. If the investigators had withdrawn participants in retrospect who did not meet the study criteria for a myocardial infarction, they would have withdrawn more participants from the intervention groups (those with no documented infarction plus those with small infarction) than from the control group (those with no infarction). This would have produced a bias in later comparisons. On the other hand, it could be assumed that a similar number of truly ineligible participants were randomized to the intervention groups and to the control group. In order to maintain comparability, the investigators might have decided to withdraw the same number of participants from each group. The ineligible participants in the control group could have been readily identified. However, the participants in the intervention groups who were truly

Table 18.1 Mortality by study group and eligibility status in the Anturane Reinfarction Trial

	Randomized	Percent mortality	Ineligible	Percent mortality	Eligible	Percent mortality
Sulfinpyrazone	813	9.1	38	26.3	775	8.3
Placebo	816	10.9	33	12.1	783	10.9

ineligible had to be distinguished from those made to appear ineligible by the effects of the interventions. This would have been difficult, if not impossible. In the Multicenter Investigation for the Limitation of Infarct Size (MILIS) for example, all randomized participants were retained in the analysis [26].

An example of possible bias because of withdrawal of ineligible participants is found in the Anturane Reinfarction Trial, which compared sulfinpyrazone with placebo in participants who had recently suffered a myocardial infarction [27–29]. As seen in Table 18.1, of 1,629 randomized participants (813 to sulfinpyrazone, 816 to placebo), 71 were subsequently found to be ineligible. Thirty-eight had been assigned to sulfinpyrazone and 33 to placebo. Despite relatively clear definitions of eligibility and comparable numbers of participants withdrawn, mortality among these ineligible participants was 26.3% in the sulfinpyrazone group (10 of 38) and 12.1% in the placebo group (4 of 33) [27]. The eligible placebo group participants had a mortality of 10.9%, similar to the 12.1% seen among the ineligible participants. In contrast, the eligible participants on sulfinpyrazone had a mortality of 8.3%, less than one-third that of the ineligible participants. Including all 1,629 participants in the analysis gave 9.1% mortality in the sulfinpyrazone group, and 10.9% mortality in the placebo group ($p = .20$). Withdrawing the 71 ineligible participants (and 14 deaths, 10 vs. 4) gave an almost significant $p = .07$.

Stimulated by criticisms of the study, the investigators initiated a reevaluation of the Anturane Reinfarction Trial results. An independent group of reviewers examined all reports of deaths in the trial [29]. Instead of 14 deceased participants who were ineligible, it found 19; 12 in the sulfinpyrazone group and seven in the placebo group. Thus, supposedly clear criteria for ineligibility can be judged differently. This trial was an early example that affirmed the value of the intention-to-treat principle.

Three trial design policies that relate to withdrawals because of entry criteria violations have been discussed by Peto et al. [12]. The first policy is not to enroll participants until all the diagnostic tests have been confirmed and all the entry criteria have been carefully checked. Once enrollment takes place, no withdrawals from the trial are allowed. For some studies, such as the one on limiting infarct size, this policy cannot be applied because firm diagnoses cannot be ascertained prior to the time when intervention has to be initiated.

The second policy is to enroll marginal or unconfirmed cases and later withdraw from analysis those participants who are proven to have been misdiagnosed. This would be allowed, however, only if the decision to withdraw is based on data

collected before enrollment. Any process of deciding upon withdrawal of a participant from a study group should be done blinded with respect to the participant's outcome and group assignment.

A third policy is to enroll some participants with unconfirmed diagnoses and to allow no withdrawals. This procedure is always valid in that the investigator compares two randomized groups which are comparable at baseline. However, this policy is conservative because each group contains some participants who might not benefit from the intervention. Thus, the overall trial may have less power to detect differences of interest.

A modification to these three policies has been recommended [22]. Every effort should be made to establish the eligibility of participants prior to any randomization. No withdrawals should be allowed and the analyses should include all participants enrolled. Analyses based on only those truly eligible may be performed. If the analyses of data from all enrolled participants and from those eligible agree, the interpretation of the results is clear, at least with respect to participant eligibility. If the results differ, however, the investigator must be very cautious in her interpretation. In general, she should emphasize the analysis with all the enrolled participants because that analysis is always valid.

Any policy on withdrawals should be stated in the study protocol before the start of the study. Though the enrolled cohort is never a random sample, in general, the desired aim is to make the recruited cohort as similar as possible to the population in which the intervention will be used in clinical practice, so withdrawal of participants from the trial or participants' data from an analysis after the decision to treat should be extremely rare. The actual decision to withdraw specific participants should be done without knowledge of the study group, ideally by someone not directly involved in the trial. Of special concern is withdrawal based on review of selected cases, particularly if the decision rests on a subjective interpretation. Even in double-blind trials, blinding may not be perfect, and the investigator may supply information for the eligibility review differentially depending upon study group and health status. Therefore, withdrawal should be done early in the course of follow-up, before a response variable has occurred, and with a minimum of data exchange between the investigator and the person making the decision to withdraw the participant. This withdrawal approach does not preclude a later challenge by readers of the report, on the basis of potential bias. It should, however, remove the concern that the withdrawal policy was dependent on the outcome of the trial. The withdrawal rules should not be based on knowledge of study outcomes. Even when these guidelines are followed, if the number of withdrawals is high, if the number of entry criteria violations is substantially different in the study groups, or if the event rates in the withdrawn participants are different between the groups, the question will certainly be raised whether bias played a role in the decision to withdraw participants.

Nonadherence

Nonadherence to the prescribed intervention or control regimen is another reason that participants are withdrawn from analysis [40–59]. One version of this is to define an "on treatment" analysis that eliminates any participant who does not adhere to the intervention by some specified amount, as defined in the protocol. One form of nonadherence may be drop-outs and drop-ins (Chap. 14). Drop-outs are participants in the intervention arm who do not adhere to the regimen. Drop-ins are participants in the control arm who receive the intervention. The decision not to adhere to the protocol intervention may be made by the participant, his primary care physician, or the trial investigator. Nonadherence may be due to adverse events in either the intervention or control group, loss of participant interest or perceived benefit, changes in the underlying condition of a participant, or a variety of other reasons.

Withdrawal from analysis of participants who do not adhere to the intervention regimens specified in the study design is often proposed. The motivation for withdrawal of nonadherent participants is that the trial is not a "fair test" of the ideal intervention with these participants included. For example, there may be a few participants in the intervention group who took little or no therapy. One might argue that if participants do not take their medication, they certainly cannot benefit from it. There could also be participants in the control group who frequently receive the study medication. The intervention and control groups are thus "contaminated." Proponents of withdrawal of nonadherent participants argue that removal of these participants keeps the trial closer to what was intended; that is, a comparison of optimal intervention versus control. The impact of nonadherence on the trial findings is that any observed benefit of the intervention, as compared to the control, will be reduced, making the trial less powerful than it planned. Newcombe [11], for example, discusses the implication of adherence for the analysis as well as the design and sample size. We discuss this in Chap. 8.

A policy of withdrawal from analysis because of participant nonadherence can lead to bias. The overwhelming reason is that participant adherence to a protocol may be related to the outcome. In other words, there may an effect of adherence on the outcome which is independent of the intervention. Certainly, if nonadherence is greater in one group than another, for example if the intervention produces many adverse events, then withdrawal of nonadherent participants could lead to bias. Even if the frequency of nonadherence is the same for the intervention and control groups, the reasons for nonadherence in each group may differ and may involve different types of participants. The concern would always be whether the same types of participants were withdrawn in the same proportion from each group or whether an imbalance had been created. Of course, an investigator could probably neither confirm nor refute the possibility of bias.

For noninferiority trials (see Chaps. 3 and 5), nonadherence may make the two interventions arms to look more alike and thus create bias towards the claim of noninferiority [13, 60]. Any attempt to use only adherers in a noninferiority trial,

Table 18.2 Percent mortality by study group and level of adherence in the Coronary Drug Project

		Drug adherence	
	Overall	≥80%	<80%
Clofibrate	18.2	15.0	24.6
Placebo	19.4	15.1	28.2

Table 18.3 Percent mortality by study group and degree of adherence in the Aspirin Myocardial Infarction Study

	Overall	Good adherence	Poor adherence
Aspirin	10.9	6.1	21.9
Placebo	9.7	5.1	22.0

though, could be biased in unknown directions, thus rendering the results uninterpretable. Again, the best policy is to design a trial to have minimum nonadherence, power the trial to overcome non-preventable nonadherence and then accept the results using the principle of intention-to-treat.

The Coronary Drug Project evaluated several lipid-lowering drugs in people several years after a myocardial infarction. In participants on one of the drugs, clofibrate, total 5-year mortality was 18.2%, as compared with 19.4% in control participants [31, 57]. Among the clofibrate participants, those who had at least 80% adherence to therapy had a mortality of 15%, whereas the poor adherers had a mortality of 24.6% (Table 18.2). This seeming benefit from taking clofibrate was, unfortunately, mirrored in the group taking placebo, 15.1% vs. 28.2%. A similar pattern (Table 18.3) was noted in the Aspirin Myocardial Infarction Study [58]. Overall, no difference in mortality was seen between the aspirin-treated group (10.9%) and the placebo-treated group (9.7%). Good adherers to aspirin had a mortality of 6.1%; poor adherers had a mortality of 21.9%. In the placebo group, the rates were 5.1% and 22%.

A trial of antibiotic prophylaxis in cancer patients also demonstrated a relationship between adherence and benefit in both the intervention and placebo groups [43]. Among the participants assigned to intervention, efficacy in reducing fever or infection was 82% in excellent adherers, 64% in good adherers, and 31% in poor adherers. Among the placebo participants, the corresponding figures were 68%, 56%, and 0%.

Another pattern is noted in a three-arm trial comparing two beta-blocking drugs, propranolol and atenolol, with placebo [59]. Approximately equal numbers of participants in each group stopped taking their medication. In the placebo group, adherers and nonadherers had similar mortality: 11.2% and 12.5%, respectively. Nonadherers to the interventions, however, had death rates several times greater than did the adherers: 15.9% to 3.4% in those on propranolol and 17.6% to 2.6% in those on atenolol. Thus, even though the numbers of nonadherers in each arm were equal, their risk characteristics as reflected by their morality rates were obviously different.

Pledger [51] provides an analogous example for a trial of schizophrenia. Participants were randomized to chlorpromazine or placebo and the 1-year relapse rates were measured. The overall comparison was a 27.8% relapse rate on active medication and 52.8% for those on placebo. The participants were categorized into low or high adherence subgroups. Among the active medication participants, the relapse rate was 61.2% for low adherence and 16.8% for high adherence. However, the relapse rate was 74.7% and 28.0% for the corresponding adherence groups on placebo.

Another example of placebo adherence versus nonadherence is reported by Oakes et al. [49]. A trial of 2,466 heart attack participants compared diltiazem with placebo over a period of 4 years with time to first cardiac event as the primary outcome. Cardiac death or all-cause mortality were additional outcome measures. The trial was initially analyzed according to intention-to-treat with no significant effect of treatment. Qualitative interaction effects were found with the presence or absence of pulmonary congestion which favored diltiazem for patients without pulmonary congestion and placebo in patients with pulmonary congestion. Interestingly, for participants without pulmonary congestion, the hazard ratio or relative risk for time to first cardiac event was 0.92 for those off placebo compared to those on placebo. For participants with pulmonary congestion, the hazard ratio was 2.86 for participants off placebo compared to those on placebo. For time to cardiac death and to all-cause mortality, hazard ratios exceeded 1.68 in both pulmonary congestion subgroups. This again suggests that placebo adherence is a powerful prognostic indicator and argues for the intention-to-treat analysis.

The definition of nonadherence can also have a major impact on the analysis. This is demonstrated by reanalysis of a trial in breast cancer patients by Redmond et al. [52]. This trial compared a complex chemotherapy regimen with placebo as adjuvant therapy following surgery with disease-free survival as the primary outcome. To illustrate the challenges of trying to adjust analyses for adherence, two measures of adherence were created. Adherence was defined as the fraction of chemotherapy taken while on the study to what was defined by the protocol as a full course. One analysis (Method I) divided participants into good adherers ($\geq$85%), moderate adherers (65–84%) and poor adherers ($<$65%). Using this definition, placebo adherers had a superior disease-free survival than moderate adherers who did better than poor adherers (Fig. 18.1). This pattern of outcome in the placebo group is similar to the CDP clofibrate example. The authors performed a second analysis (Method II) changing the definition of adherence slightly. In this case, adherence was defined as the fraction of chemotherapy taken while on study to what should have been taken while still on study before being taken off treatment for some reason. Note that the previous definition (Method I) compared chemotherapy taken to what would have been taken had the participant survived to the end and adhered perfectly. This subtle difference in definition changed the order of outcome in the placebo group. Here, the poor placebo adherers had the best disease-free survival and the best adherers had a disease-free survival in-between the moderate and poor adherers. Of special importance is that the participants in this example were all on placebo. Thus, adherence is itself an outcome and trying to adjust one

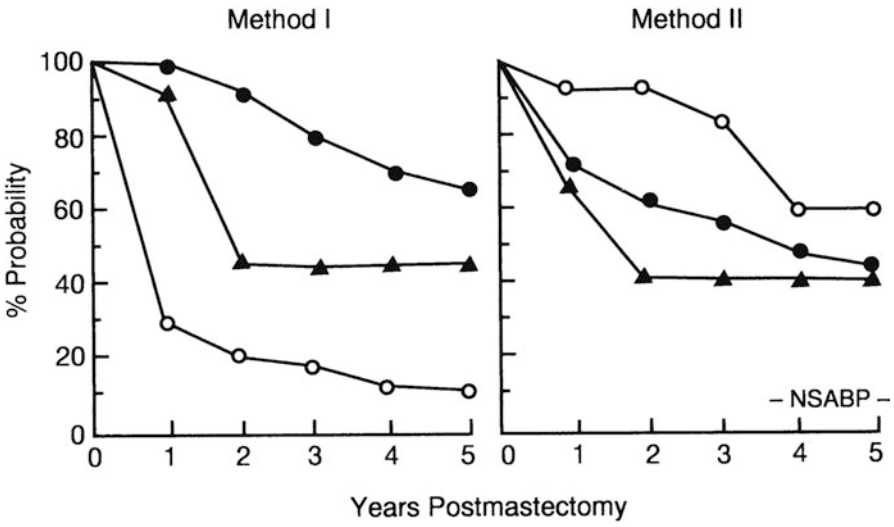

Fig. 18.1 Percentage of disease free survival related to adherence levels of placebo; methods I and II definition of compliance in National Surgical Adjuvant Breast Program (NSABP). Three levels of compliance are: Good (>85%) Moderate (65–84%) Poor (<65%) [52]

outcome (the primary response variable) for another outcome (adherence) can lead to misleading results.

Detre and Peduzzi have argued that, although as a general rule nonadherent participants should be analyzed according to the study group to which they were assigned, there can be exceptions. They presented an example from the VA coronary bypass surgery trial [40]. In that trial, a number of participants assigned to medical intervention crossed over to surgery. Contrary to expectation, these participants were at similar risk of having an event, after adjusting for a variety of baseline factors, as those who did not crossover. Therefore, the authors argued that the non-adherers should be kept in their original groups, but can be censored at the time of crossover. This may be true, but, as seen in the Coronary Drug Project [57], adjustment for known variables does not always account for the observed response. The differences in mortality between adherers and nonadherers remained even after adjustment. Thus, other unknown or unmeasured variables were of critical importance.

Some might think that if rules for withdrawing participants are specified in advance, withdrawals for nonadherence are legitimate. However, the potential for bias cannot be avoided simply because the investigator states, ahead of time, the intention to withdraw participants. This is true even if the investigator is blinded to the group assignment of a participant at the time of withdrawal. Participants were not withdrawn from the analyses in the above examples. However, had a rule allowing withdrawal of participants with poor adherence been specified in advance, the results described above would have been obtained. The type of participants withdrawn would have been different in the intervention and control groups and

would have resulted in the analysis of non-comparable groups of adherers. Unfortunately, as noted, the patterns of possible bias can vary and depend on the precise definition of adherence. Neither the magnitude nor direction of that bias is easily assessed or compensated for in analysis.

Adherence is also a response to the intervention. If participant adherence to an intervention is poor compared to that of participants in the control group, widespread use of this therapy in clinical practice may not be reasonably expected. An intervention may be effective, but may be of little value if it cannot be tolerated by a large portion of the participants. For example, in the Coronary Drug Project, the niacin arm showed a favorable trend for mortality over 7 years, compared with placebo, but niacin caused "hot flashes" and was not easily tolerated [31]. The development of slow release formulations that reduce pharmacologic peaks has lessened the occurrence of side effects.

It is therefore recommended that no participants be withdrawn from analysis in superiority trials for lack of adherence. The price an investigator must pay for this policy is possibly reduced power because some participants who are included in the analysis may not be on optimal intervention. For limited or moderate nonadherence, one can compensate by increasing the sample size, as discussed in Chap. 8, although doing so is costly.

Missing or Poor Quality Data

In most trials, participants have data missing for a variety of reasons. Perhaps they were not able to keep their scheduled clinic visits or were unable to perform or undergo the particular procedures or assessments. In some cases, follow-up of the participant was not completed as outlined in the protocol. The challenge is how to deal with missing data or data of such poor quality that they are in essence missing. One approach is to withdraw participants who have poor data completely from the analysis [26, 61, 62]. However, the remaining subset may no longer be representative of the population randomized and there is no guarantee that the validity of the randomization has been maintained in this process.

There is a vast literature on approaches to dealing with missing data [63–73]. Many of these methods assume that the data are missing at random; that is, the probability of a measurement not being observed does not depend on what its value would have been. In some contexts, this may be a reasonable assumption, but for clinical trials, and clinical research in general, it would be difficult to confirm. It is, in fact, probably not a valid assumption, as the reason the data are missing is often associated with the health status of the participant. Thus, during trial design and conduct, every effort must be made to minimize missing data. If the amount of missing data is relatively small, then the available analytic methods will probably be helpful. If the amount of missing data is substantial, there may be no method capable of rescuing the trial. In this section, we discuss some of the issues that must be kept in mind when analyzing a trial with missing data.

Rubin [72] provided a definition of missing data mechanisms. If data are missing for reasons unrelated to the measurement that would have been observed and unrelated to covariates, then the data are "missing completely at random." Statistical analyses based on likelihood inference are valid when the data are missing at random or missing completely at random. If a measure or index allows a researcher to estimate the probability of having missing data, say in a participant with poor adherence to the protocol, then using methods proposed by Rubin and others might allow some adjustment to reduce bias [66, 71, 72, 74]. However, adherence, as indicated earlier, is often associated with a participant's outcome and attempts to adjust for adherence can lead to misleading results.

If participants do not adhere to the intervention and also do not return for follow-up visits, the primary outcome measured may not be obtained unless it is survival or some easily ascertained event. In this situation, an intention-to-treat analysis is not feasible and no analysis is fully satisfactory. Because withdrawal of participants from the analysis is known to be problematic, one approach is to "impute" or fill in the missing data such that standard analyses can be conducted. This is appealing if the imputation process can be done without introducing bias. There are many procedures for imputation. Those based on multiple imputations are more robust than single imputation [75].

A commonly used single imputation method is to carry the last observed value forward. This method, also known as an endpoint analysis, requires the very strong and unverifiable assumption that all future observations, if they were available, would remain constant [51]. Although commonly used, the last observation carried forward method is not generally recommended [71, 73]. Using the average value for all participants with available data, or using a regression model to predict the missing value are alternatives, but in either case, the requirement that the data be missing at random is necessary for proper inference.

A more complex approach is to conduct multiple imputations, typically using regression methods, and then perform a standard analysis for each imputation. The final analysis should take into consideration the variability across the imputations. As with single imputation, the inference based on multiple imputation depends on the assumption that the data are missing at random. Other technical approaches are not described here, but in the context of a clinical trial, none is likely to be satisfactory.

Various other methods for imputing missing values have been described [63–73, 75]. Examples of some of these methods are given by Espeland et al. for a trial measuring carotid artery thickness at multiple anatomical sites using ultrasound [61]. In diagnostic procedures of this type, typically not all measurements can be made. Several imputation methods, based on a mixed effects linear model where regression coefficient and a covariance structure (i.e., variances and correlations), were estimated. Once these were known, this regression equation was the basis for the imputation. Several imputation strategies were used based on different methods of estimating the parameters and whether treatment differences were assumed or not. Most of the imputation strategies gave similar results when the trial data were

analyzed. The results indicated up to a 20% increase in efficiency compared to using available data in cross sectional averages.

For repeated measures, imputation techniques such as these are useful if the data are missing at random; that is, the probability of missing data is not dependent on the measurement that would have been observed or on the preceding measurements. Unfortunately, it is unlikely that data are missing at random. The best that can be offered, therefore, is a series of analyses, each exploring different approaches to the imputation problem. If all, or most, are in general agreement qualitatively, then the results are more persuasive. All analyses should be presented, not just the one with the preferred results.

In long-term trials participants may be lost to follow-up or refuse to continue their participation. In this situation, the status of the participant with regard to any response variable cannot be determined. If mortality is the primary response variable and if the participant fails to return to the clinic, his survival status may still be obtained. If a death has occurred, the date of death can be ascertained. In the Coronary Drug Project [31] where survival experience over 60 months was the primary response variable, four of 5,011 participants were lost to follow-up (one in a placebo group, three in one treatment group, and none in another treatment group). The Lipid Research Clinics Coronary Primary Prevention Trial [47] followed over 3,800 participants for an average of 7.4 years, and was able to assess vital status on all. The Physicians' Health study of over 20,000 US male physicians had complete follow-up for survival status [76]. Many other large simple trials, such as GUSTO [39], have similar nearly complete follow-up experience. Obtaining such low loss to follow-up rates, however, required special effort. In the Women's Health Initiative (WHI), one portion evaluated the possible benefits of hormone replacement therapy (estrogen plus progestin) compared with placebo in post-menopausal women. Of the 16,025 participants, 3.5% were lost to follow-up and did not provide 18 month data [77].

For some conditions, e.g., trials of treatment for substance abuse, many participants fail to return for follow-up visits, and missing data can be 25–30% or even more. Efforts to account for missing data must be made, recognizing that biases may very well exist.

An investigator may not be able to obtain any information on some kinds of response variables. For example, if a participant is to have blood pressure measured at the last follow-up visit 12 months after randomization and the participant does not show up for that visit, this blood pressure can never be retrieved. Even if the participant is contacted later, the later measurement does not truly represent the 12-month blood pressure. In some situations, substitutions may be permitted, but, in general, this will not be a satisfactory solution. An investigator needs to make every effort to have participants come in for their scheduled visits in order to keep losses to follow-up at a minimum. In the Intermittent Positive Pressure Breathing (IPPB) trial, repeated pulmonary function measurements were required for participants with chronic obstructive pulmonary disease [62]. However, some participants who had deteriorated could not perform the required test. A similar problem existed for

the Multicenter Investigation of the Limitation of Infarct Size (MILIS) where infarct size could not be obtained in many of the sickest participants [26].

Individuals with chronic obstructive pulmonary disease typically decline in their pulmonary function and this decline may lead to death, as happened to some participants in the IPPB trial. In this case, the missing data were not missing at random and censoring was said to be informative. One simple method for cases such as the IPPB study is to define a decreased performance level considered to be a clinical event. Then the analysis can be based on time to the clinical event of deterioration or death, incorporating both pieces of information. Survival analysis, though, assumes that loss of follow-up is random and independent of risk of the event. Methods relaxing the missing at random assumption have been proposed [78, 79], but require other strong assumptions, the details of which are beyond the scope of this text.

If the number of participants lost to follow-up differs in the study groups, the analysis of the data could be biased. For example, participants who are taking a new drug that has adverse effects may, as a consequence, miss scheduled clinic visits. Events may occur but be unobserved. These losses to follow-up would probably not be the same in the control group. In this situation, there may be a bias favoring the new drug. Even if the number lost to follow-up is the same in each study group, the possibility of bias still exists because the participants who are lost in one group may have quite different prognoses and outcomes than those in the other group.

An example of differential follow-up was reported by the Comparison of Medical Therapy, Pacing, and Defibrillation in Chronic Heart Failure (COMPANION) trial [80]. COMPANION compared a cardiac pacemaker or a pacemaker plus defibrillator with best pharmacologic treatment in people with chronic heart failure. Over 1,500 participants were randomized. Two primary outcomes were assessed; death and death plus hospitalization. Individuals randomized to one of the device arms did not know to which device they had been assigned, but those on the pharmacologic treatment arm were aware that no device had been installed. During the course of the trial, the pacemaker plus defibrillator devices, made by two different manufacturers, were approved by a regulatory agency. As a result, participants in the pharmacologic treatment arm began to drop-out from the trial and some also withdrew their consent. Many requested one of the newly approved devices. Thus, when the trial was nearing completion, the withdrawal rate was 26% in the pharmacologic treatment arm and 6–7% in the device arms. Additionally, no further follow-up information could be collected on those who withdrew consent. Clearly, censoring at the time of withdrawal was not random and the possibility that it was related to disease status could not be ruled out, thus creating the possibility of serious bias. This situation could have jeopardized an otherwise well designed and conducted trial in people with a serious medical condition. However, the investigators initiated a complicated process of reconsenting the participants to allow for collection of the primary outcomes. After completing this process, assessment of the status for death plus hospitalization and vital status were 91% and 96%, respectively, in the pharmacologic treatment group. Outcome ascertainment for the two device arms was 99% or better. The final results demonstrated that both the

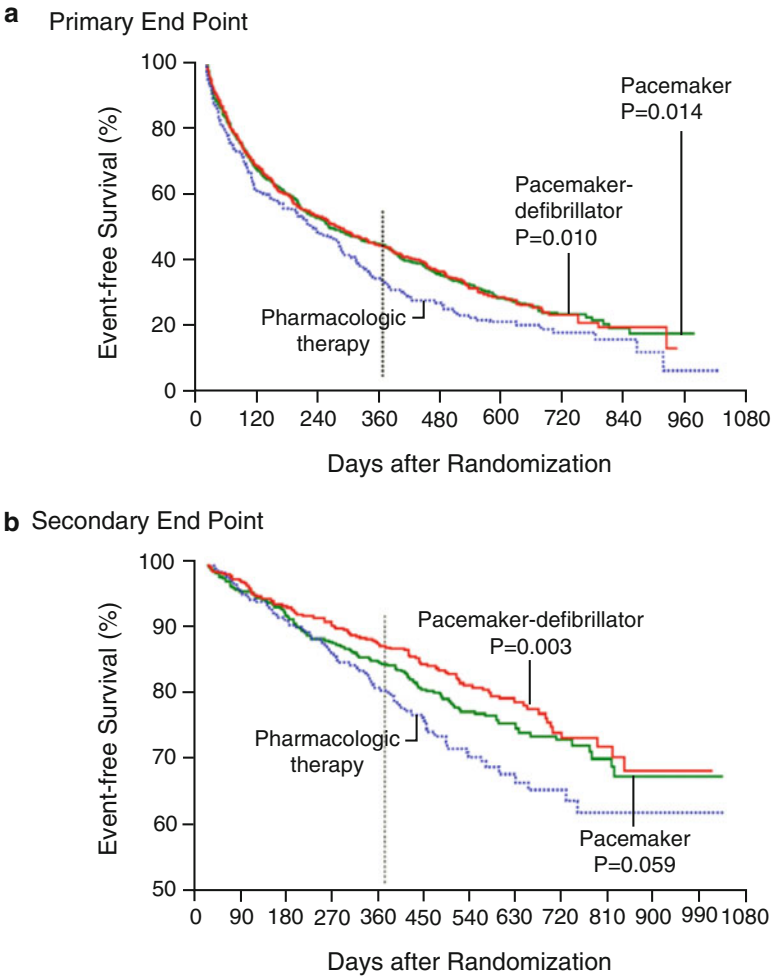

Fig. 18.2 Kaplan-Meier estimates in COMPANION trial. (**a**) The time to primary end point of death from hospitalization for any cause. (**b**) The time to the secondary end point of death from any cause [80]

pacemaker and the defibrillator plus pacemaker reduced death plus hospitalization and further that the defibrillator plus pacemaker reduced mortality. These results were important in the treatment of chronic heart failure. However, not correcting for the initial differential loss to follow-up would have rendered the COMPANION trial data perhaps uninterpretable. In Fig. 18.2, the Kaplan Meier curves for mortality for the two intervention arms are provided with the most complete data available.

Often, protocol designs call for follow-up to terminate at some period, for example 7, 14, or 30 days, after a participant has stopped adhering to his or her

intervention, even though the intended duration of intervention would not have ended. The concept is that "off intervention" means "off study"; i.e., assessment for nonadherent participants ends when intervention ends. We do not endorse this concept. Although time to event analysis may be censored at the time of last follow-up, going off intervention or control is not likely random and may be related to participant health status. Important events, including serious adverse events, may occur beyond the follow-up period and might be related to the intervention. As noted above, though, survival analysis assumes that censoring is independent of the primary event. The practice of not counting events at the time of, or shortly after, intervention discontinuation is all too common, and can lead to problems in the interpretation of the final results. An instructive example is the Adenomatous Polyp Prevention on Vioxx (APPROVe) trial [81]. This randomized double blind trial compared a cyclo-oxygenase (COX)-II inhibitor with placebo in people with a history of colorectal adenomas. Previous trials of COX-II inhibitors had raised concern regarding long term cardiovascular risk. Thus, while the APPROVe trial was a cancer prevention trial, attention also focused on cardiovascular events, in particular thrombotic events and cardiovascular death, nonfatal myocardial infarction, and nonfatal stroke. However, participants who stopped taking their medication during the trial were not followed beyond 14 days after the time of discontinuation. The Kaplan-Meier cardiovascular risk curve is shown in Fig. 18.3. Note that for the first 18 months the two curves are similar and then begin to diverge. Controversy arose as to whether there was an 18-month lag time in the occurrence of cardiovascular events for this particular COX-II inhibitor [82, 83].

Due to the controversy, the investigators and sponsor launched an effort to collect information on all trial participants for at least a year after stopping study treatment. This extended follow-up, referred to here as APPROVe + 1, was able to collect selected cardiovascular events of nonfatal myocardial infarction, nonfatal stroke and cardiovascular death [84], as shown in Fig. 18.4. The time to event curves begin to separate from the beginning and continue throughout the extended follow-up, with a hazard ratio of 1.8 ($p = 0.006$). There was a corresponding statistically nonsignificant increase in mortality.

Censoring follow-up when participants go off their intervention is a common error that leads to problems like those encountered by the APPROVe trial. Going off intervention, and thus censoring follow-up at some number of days afterwards, is not likely to be independent of the disease process or how a participant is doing. At least, it would be difficult to demonstrate such independence. Yet, survival analysis and most other analyses assume that the censoring is independent. The principle lesson here is that "off intervention should not mean off study."

An outlier is an extreme value significantly different from the remaining values. The concern is whether extreme values in the sample should be included in the analysis. This question may apply to a laboratory result, to the data from one of several areas in a hospital or from a clinic in a multicenter trial. Removing outliers is not recommended unless the data can be clearly shown to be erroneous. Even though a value may be an outlier, it could be correct, indicating that on occasions an extreme result is possible. This fact could be very important and should not be

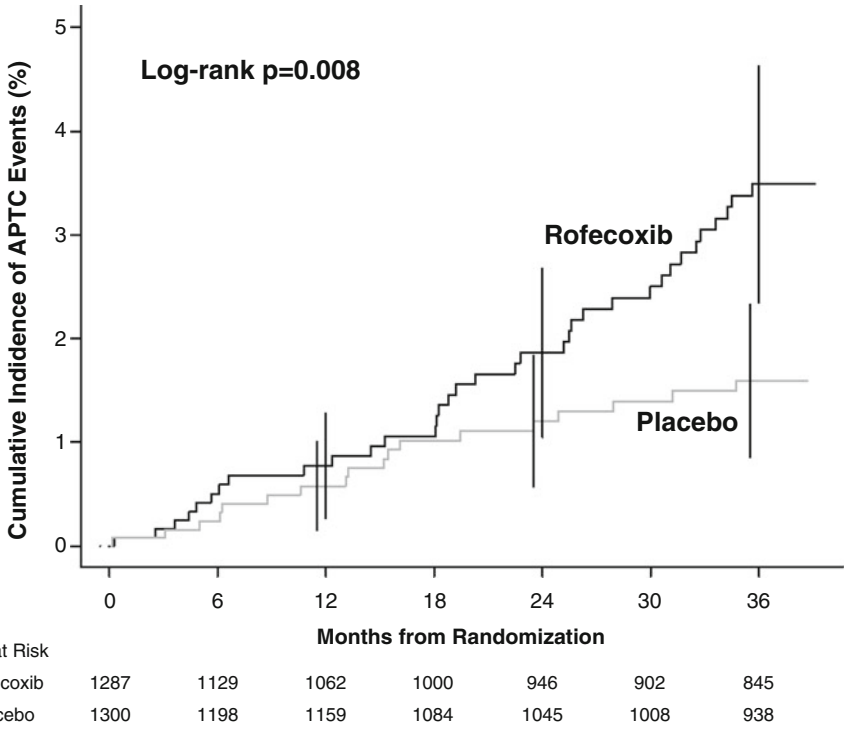

Fig. 18.3 APPROVe—Kaplan-Meier estimates of time to death from the AntiPlatelet Trialists' Collaborative (APTC) outcomes (cardiovascular causes, nonfatal myocardial infarction of nonfatal stroke) with censoring 14 days after participants stopped therapy [84]. Reproduced with the permission of Elsevier Ltd. for *Lancet*

ignored. Long ago Kruskall [85] suggested carrying out an analysis with and without the "wild observation." If the conclusions vary depending on whether the outlier values are included or excluded, one should view any conclusions cautiously. Procedures for detecting extreme observations have been discussed [86–89], and the publications cited can be consulted for further detail.

An interesting example given by Canner et al. [86] concerns the Coronary Drug Project. The authors plotted the distributions of four response variables for each of the 53 clinics in that multicenter trial. Using total mortality as the response variable, no clinics were outlying. When nonfatal myocardial infarction was the outcome, only one clinic was an outlier. With congestive heart failure and angina pectoris, response variables which are probably less well defined, there were nine and eight outlying clinics, respectively.

In conclusion, missing data can create problems. Though methods which allow for missing data exist, they require certain assumptions which are not likely to be true. Every attempt should made to minimize missing data, and investigators should be aware of the potential for bias

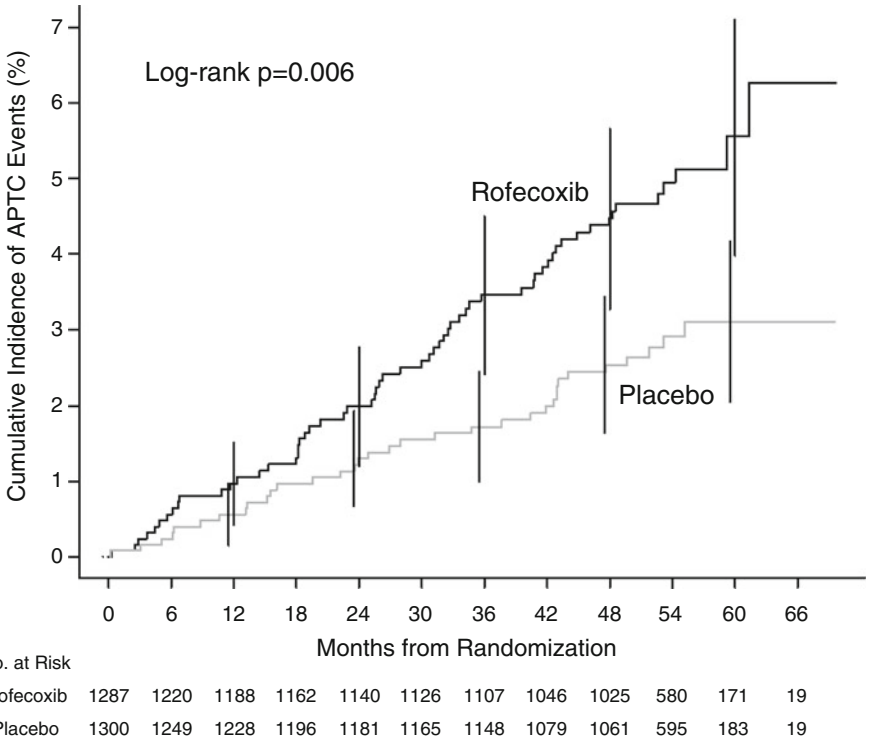

Fig. 18.4 APPROVe Kaplan-Meier estimates of time to death for the AntiPlatelet Trialists' Collaborative (APTC) outcome (cardiovascular causes, nonfatal myocardial infarction or nonfatal stroke) counting all events observed for an additional year of follow-up after trial was initially terminated [84]. Reproduced with the permission

Competing Events

Competing events are those that preclude the assessment of the primary response variable. They can reduce the power of the trial by decreasing the number of participants available for follow-up. If the intervention can affect the competing event, there is also the risk of bias. In some clinical trials, the primary response variable may be cause-specific mortality, such as death due to myocardial infarction or sudden death, rather than total mortality [90–93]. The reason for using cause-specific death as a response variable is that a therapy often has specific mechanisms of action that may be effective against a disease or condition. In this situation, measuring death from all causes, at least some of which are not likely to be affected by the intervention, can "dilute" the results. For example, a study drug may be anti-arrhythmic and thus sudden cardiac death might be the selected response variable. Other causes of death such as cancer and accidents would be competing events.

Even if the response variable is not cause-specific mortality, death may be a factor in the analysis. This is particularly an issue in long term trials in the elderly or high risk populations. If a participant dies, future measurements will be missing. Analysis of nonfatal events in surviving participants has the potential for bias, especially if the mortality rates are different in the two groups.

In a study in which cause-specific mortality is the primary response variable, deaths from other causes are treated statistically as though the participants were lost to follow-up from the time of death (Chap. 15) and these deaths are not counted in the analysis. In this situation, the analysis, however, must go beyond merely examining the primary response variable. An intervention may or may not be effective in treating the condition of interest but could be harmful in other respects. Therefore, total mortality should be considered as well as cause-specific fatal events. Similar considerations need to be made when death occurs in studies using nonfatal primary response variables. This can be done by considering tables that show the number of times the individual events occur, one such event per person. No completely satisfactory solution exists for handling competing events. At the very least, the investigator should report all major outcome categories; for example, total mortality, as well as cause-specific mortality and morbid events.

In many cases, there may be recurring events. Many trials simply evaluate the time to the first event and do not count the additional events in the time to event analysis. Tables may show the total number of events in each intervention arm. Some attempts to further analyze recurrent events have been made, using for example the data from the COMPANION trial [80, 90]. Software exists for these methods [94, 95]; however, the technical details of these methods are complicated and will not be covered in this text.

Composite Outcomes

In recent years, many trials have used combinations of clinical and other outcomes as a composite response variable [90–93]. One major motivation is to increase the event rate and thus reduce the sample size that might otherwise have been required had just one of the components been selected as the primary outcome. Another motivation is to combine events that have a presumed common etiology and thus get an overall estimate of effect. The sample size is usually not based on any single component.

There are challenges in using a composite outcome [96, 97]. The components may not have equal weight or clinical importance, especially as softer outcomes are added. The components may go in opposite directions or at least not be consistent in indicating intervention effect. One component may dominate the composite. Results with any single component are based on a smaller number of events and thus the power for that component is greatly reduced. Rarely do we find significance for a component, nor should we expect it in general. Regardless of the composition of the composite, analyses should be conducted for each component, or in some

cascading sequence. For example, if the composite were death, myocardial infarction, stroke or heart failure hospitalization, the analysis sequence might be death, death plus myocardial infarction or stroke, and death plus heart failure hospitalization. The reason for including death is to take into account competing risk of death for the other components, in addition to its obvious clinical importance.

As pointed out in Chap. 3, it is essential that follow-up continue after the first event in the composite outcome occurs. Analysis will include looking at the contribution of each component to the overall but should also include time to event for each component separately. As indicated, if follow-up does not continue, only partial results are available for each component and analysis of those events separately could be misleading.

There are several examples where the use of a composite such as death, myocardial infarction and stroke has been used as a primary or leading secondary outcome [34, 36–38]. These outcomes are all clinically relevant. In these trials, the outcomes all trended in the same direction. However, that is not always the case.

In the Pravastatin or Atorvastatin Evaluation and Infection Therapy (PROVE IT) trial, the 80 mg atorvastatin arm was more effective than the 40 mg pravastatin arm in reducing the incidence of death, myocardial infarction, stroke, required hospitalization due to unstable angina and revascularization [91]. Stroke results, one of the key components, went in the opposite direction. These results complicate the interpretation. Should investigators think that the atorvastatin improves the composite or just those components that are in the same direction as the composite? As would be expected, the differences for the components were not, in themselves, statistically significant.

Another interesting example is provided by the Women's Health Initiative (WHI) which was a large factorial design trial post-menopausal women [77]. As discussed earlier and in Chap. 16, one part involved hormone replacement therapy which contained two strata, women with a uterus and those without. Women with a uterus received estrogen plus progestin or matching placebo; those without a uterus received estrogen alone or a matching placebo. Due to the multiple actions of hormone replacement therapy, one response variable was a global outcome mortality, coronary heart disease, bone loss reflected by hip fracture rates, breast cancer, colorectal cancer, pulmonary embolism, and stroke. As seen in Fig. 16.7, for the estrogen plus progestin stratum, there was essentially no effect on mortality and a small but nonsignificant effect in the global index, when compared to placebo. However, as shown in Fig. 16.6, the various components went in different directions. Hip fracture and colorectal cancer had a favorable response to hormone replacement therapy. Pulmonary embolism, stroke and coronary heart disease went in an unfavorable direction. Thus, any interpretation of the global index, which is a composite, requires careful examination of the components. Of course, few trials would have been designed with adequate power for the individual components so the interpretation must be qualitative, looking for consistency and biological plausibility.

The Look AHEAD trial examined whether a long-term lifestyle intervention for weight loss would decrease cardiovascular morbidity and mortality in overweight

or obese patients with type 2 diabetes [98]. The primary outcome was a composite of death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke. During follow-up, the Data and Safety Monitoring Board (DSMB) alerted the investigators that the event rate for the primary outcome was dramatically lower than expected, less than a third [99]. The protocol was changed to include hospitalization for angina as a way of increasing the event rate, and this turned out to be the most frequent component in the revised composite, which had an incidence about 50% higher than the original composite. Unfortunately, hospitalization for angina showed markedly less effect of the intervention [100]. Using the original composite would not have changed the trial's outcome, which was negative, but this experience underscores the importance of giving full consideration of a candidate component's likely response to the intervention, as well as to its incidence rate.

Experience suggests that composite outcome variables should be used cautiously and only include those components that have relatively equal clinical importance, frequency, and anticipated response to the presumed mechanism of action of the intervention [96]. As softer and less relevant outcomes are added, the interpretation becomes less clear, particularly if the less important component occurs more frequently than others, driving the overall result. Significance by any individual component cannot be expected but there should be a plausible consistency across the components.

Covariate Adjustment

The goal in a clinical trial is to have study groups that are comparable except for the intervention being studied. Even if randomization is used, all of the prognostic factors may not be perfectly balanced, especially in smaller studies. Even if no prognostic factors are significantly imbalanced in the statistical sense, an investigator may, nevertheless, observe that one or more factors favor one of the groups. In either case, covariate adjustment can be used in the analysis to minimize the effect of the differences. However, covariate adjustment is not likely to eliminate the effect of these differences. Covariance analysis for clinical trials has been reviewed in numerous articles [101–122].

Adjustment also reduces the variance in the test statistic. If the covariates are highly correlated with outcome, this can produce more sensitive analyses. The specific adjustment procedure depends on the type of covariate being adjusted for and the type of response variable being analyzed. If a covariate is discrete, or if a continuous variable is converted into intervals and made discrete, the analysis is sometimes referred to as "stratified." A *stratified analysis*, in general terms, means that the study participants are subdivided into smaller, more homogeneous groups, or strata. A comparison of study groups is made within each stratum and then averaged over all strata to achieve a summary result for the response variable. This result is adjusted for group imbalances in the discrete covariates. If a response

variable is discrete, such as the occurrence of an event, the stratified analysis might take the form of a Mantel-Haenszel statistic.

If the response variable is continuous, the stratified analysis is referred to as *analysis of covariance*. This uses a model which, typically, is linear in the covariates. A simple example for a response Y and covariate X would be $Y = \alpha_j + \beta(X - \mu) + error$ where β is a coefficient representing the importance of the covariate X and is assumed to be the same in each group, μ is the mean value of X, and α_j is a parameter for the contribution of the overall response variable jth group (e.g., $j = 1$ or 2). The basic idea is to adjust the response variable Y for any differences in the covariate X between the two groups. Under appropriate assumptions, the advantage of this method is that the continuous covariate X does not have to be divided into categories. Further details can be found in statistics textbooks [1–6, 8, 123]. If time to an event is the primary response variable, then survival analysis methods that allow for adjustments of discrete or continuous covariates may be used [106]. However, whenever models are employed, the investigator must be careful to evaluate the assumptions required and how closely they are met. Analysis of covariance can be attractive, but may be abused if linearity is assumed when the data are nonlinear, if the response curve is not parallel in each group, or if assumptions of normality are not met [122]. If measurement errors in covariates are substantial, the lack of precision can be increased [112]. For all of these reasons, covariate adjustment models may be useful in the interpretation of data, but should not be viewed as absolutely correct.

Regardless of the adjustment procedure, covariates should be measured at baseline. Except for certain factors such as age, sex, or race, any variables that are evaluated after initiation of intervention should be considered as response variables. Group comparisons of the primary response variable, adjusted for other response variables, are discouraged. Interpretation of such analyses is difficult because group comparability may be lost.

Surrogates as a Covariate

Adjustment for various surrogate outcomes may be proposed. In a trial of clofibrate [101], the authors reported that those participants who had the largest reduction in serum cholesterol had the greatest clinical improvement. However, reduction in cholesterol is probably highly correlated with adherence to the intervention regimen. Since, as discussed earlier, adherers in one group may be different from adherers in another group, analyses that adjust for a surrogate for adherence can be biased. This issue was addressed in the Coronary Drug Project [56]. Adjusted for baseline factors, the 5-year mortality was 18.8% in the clofibrate group (N = 997) and 20.2% in the placebo group (N = 2,535), an insignificant difference. For participants with baseline serum cholesterol greater than or equal to 250 mg/dl, the mortality was 17.5% and 20.6% in the clofibrate and placebo groups, respectively. No difference in mortality between the groups was noted for participants

Table 18.4 Percent 5-year mortality in the clofibrate group, by baseline cholesterol and change in cholesterol in the Coronary Drug Project's

	Baseline cholesterol	
	<250 mg/dl	≥250 mg/dl
Total	20.0	17.5
Fall in cholesterol	16.0	18.1
Rise in cholesterol	25.5	15.5

with baseline cholesterol of less than 250 mg/dl (20.0% vs. 19.9%). Those participants with lower baseline cholesterol in the clofibrate group who had a reduction in cholesterol during the trial had 16.0% mortality, as opposed to 25.5% mortality for those with a rise in cholesterol (Table 18.4). This would fit the hypothesis that lowering cholesterol is beneficial. However, in those participants with high baseline cholesterol, the situation was reversed. An 18.1% mortality was seen in those who had a fall in cholesterol, and a 15.5% mortality was noted in those who had a rise in cholesterol. The best outcome, therefore, appeared to be in participants on clofibrate whose low baseline cholesterol dropped or whose high baseline cholesterol increased. As seen earlier, adherence can affect outcomes in unexpected ways, and the same is true of surrogates for adherence.

Modeling the impact of adherence on a risk factor and thus on the response has also received attention [109, 115]. Regression models have been proposed that attempt to adjust outcome for the amount of risk factor change that could have been attained with optimum adherence. One example of this was suggested by Efron and Feldman [109] for a lipid research study. However, Albert and DeMets [115] showed that these models are very sensitive to assumptions about the independence of adherence and health status or response. If these assumptions using these regression models are violated, misleading results emerge, such as that for the clofibrate and serum cholesterol example described above.

Clinical trials of cancer treatment commonly analyze results by comparing responders to nonresponders [104, 108]. That is, those who go into remission or have a reduction in tumor size are compared to those who do not. One early survey indicated that such analyses were done in at least 20% of published reports [122]. The authors of that survey argued that statistical problems, due to lack of random assignment, and methodological problems, due both to classification of response and inherent differences between responders and nonresponders, can occur. These will often yield misleading results, as shown by Anderson et al. [104]. They pointed out that participants "who eventually become responders must survive long enough to be evaluated as responders." This factor can invalidate some statistical tests comparing responders to nonresponders. Those authors present two statistical tests that avoid bias. They note, though, that even if the tests indicate a significant difference in survival between responders and nonresponders, it cannot be concluded that increased survival is due to tumor response. Thus, aggressive intervention, which may be associated with better response, cannot be assumed to be better than less intensive intervention, which may be associated with poorer response. Anderson and colleagues state that only a truly randomized

comparison can say which intervention method is preferable. What is unsaid, and illustrated by the Coronary Drug Project examples, is that even comparison of good responders in the intervention group with good responders in the control can be misleading, because the reasons for good response may be different.

Morgan [48] provided a related example of comparing duration of response in cancer patients, where duration of response is the time from a favorable response such as tumor regression (partial or total) to remission. This is another form of defining a subgroup of post-treatment outcome, that is, tumor response. In a trial comparing two complex chemotherapy regiments (A vs. B) in small cell lung cancer, the tumor response rates were 64% and 85%, with median duration of 245 days and 262 days respectively. When only responders were analyzed, the slight imbalance in prognostic factors was substantially increased. Extensive disease was evident at baseline in 48% of one and 21% of the other treatment responder groups. Thus, while it may be theoretically possible to adjust for prognostic factors, in practice, such adjustment may decrease bias, but will not eliminate it. Because not all prognostic factors are known, any model is only an approximation to the true relationship.

The Cox proportional hazards regression model for the analysis of survival data (Chap. 15) allows for covariates in the regression to vary with time [116]. This has been suggested as a way to adjust for factors such as adherence and level of response. It should be pointed out that, like simple regression models, this approach is vulnerable to the same biases described earlier in this chapter. For example, if cholesterol level and cholesterol reduction in the CDP example were used as time dependent covariates in the Cox model, the estimator of treatment effect would be biased due to the effects shown in Table 18.4.

Rosenbaum [121] provides a nice overview of adjustment for concomitant variables that have been affected by treatment in both observational and randomized studies. He states that "adjustments for post-treatment concomitant variables should be avoided when estimating treatment effects except in rather special circumstances, since adjustments themselves can introduce a bias where none existed previously."

A number of additional methodologic attempts to adjust for adherence have also been conducted. Newcombe [11], for example, suggested adjusting estimates of intervention effect on the extent of nonadherence. Robins and Tsiatis [110] proposed a causal inference model. Lagakos et al. [46] evaluated censoring survival time, or time to an event, at the point when treatment is terminated. The rationale is that participants are no longer able to completely benefit from the therapy. However, the hazard ratio estimated by this approach is not the hazard that would have been estimated if participants had not terminated treatment. The authors stated that it is not appropriate to evaluate treatment benefit by comparing the hazard rates estimated by censoring for treatment termination [46]. Models for causal interference have also been used to explore the effects of adherence in clinical trials [124–127]. Though promising, these approaches require strong assumptions usually either known to be untrue or difficult to validate and so are not recommended as part of a primary analysis.

Baseline Variables as Covariates

The issue of stratification was first raised in the discussion of randomization (Chap. 6). For large studies, the recommendation was that stratified randomization is usually unnecessary because overall balance would nearly always be achieved and that stratification would be possible in the analysis. For smaller studies, baseline adaptive methods could be considered but the analysis should include the covariates used in the randomization. In a strict sense, analysis should always be stratified if stratification was used in the randomization. In such cases, the adjusted analysis should include not only those covariates found to be different between the groups, but also those stratified during randomization. Of course, if no stratification is done at randomization, the final analysis is less complicated since it would involve only those covariates that are imbalanced at baseline or to be of special interest associated with the outcomes.

As stated in Chap. 6, randomization tends to produce comparable groups for both measured and unmeasured baseline covariates. However, not all baseline covariates will be closely matched. Adjusting treatment effect for these baseline disparities continues to be debated. Canner [111] describes two points of view, one which argues that "if done at all, analyses should probably be limited to covariates for which there is a disparity between the treatment groups and that the unadjusted measure is to be preferred." The other view is "to adjust on only a few factors that were known from previous experience to be predictive." Canner [111], as well as Beach and Meier [107], indicate that even for moderate disparity in baseline comparability, or even if the covariates are moderately predictive, it is possible for covariate adjustment to have a nontrivial impact on the measure of treatment effect. However, Canner [111] also points out that it is "often possible to select specific covariates out of a large set in order to achieve a desired result." In addition, he shows that this issue is true for both small and large studies. For this reason, it is critical that the process for selecting covariates be specified in the protocol and adhered to in the primary analyses. Other adjustments may be used in exploratory analyses.

Another issue is testing for *covariate interaction* in a clinical trial [105, 113, 114, 118, 119]. Treatment-covariate interaction is defined when the response to treatment varies according to the value of the covariate [105]. Peto [118] defines treatment covariate interactions as quantitative or qualitative. Quantitative interactions indicate that the magnitude of treatment effect varies with the covariate but still favors the same intervention (Fig. 18.5a). Qualitative interaction involves a favorable intervention effects for some values of the covariate and unfavorable effects for others (Fig. 18.5b). A quantitative interaction, for example, would be if the benefit of treatment for blood pressure on mortality varied in degree by the level of baseline blood pressure but still favoring the same intervention (See Fig. 18.5a). A qualitative interaction would exist if lowering blood pressure was beneficial for severe hypertension, but less beneficial or even harmful for mild hypertension. Intervention effects can vary by chance across levels of the covariate, even

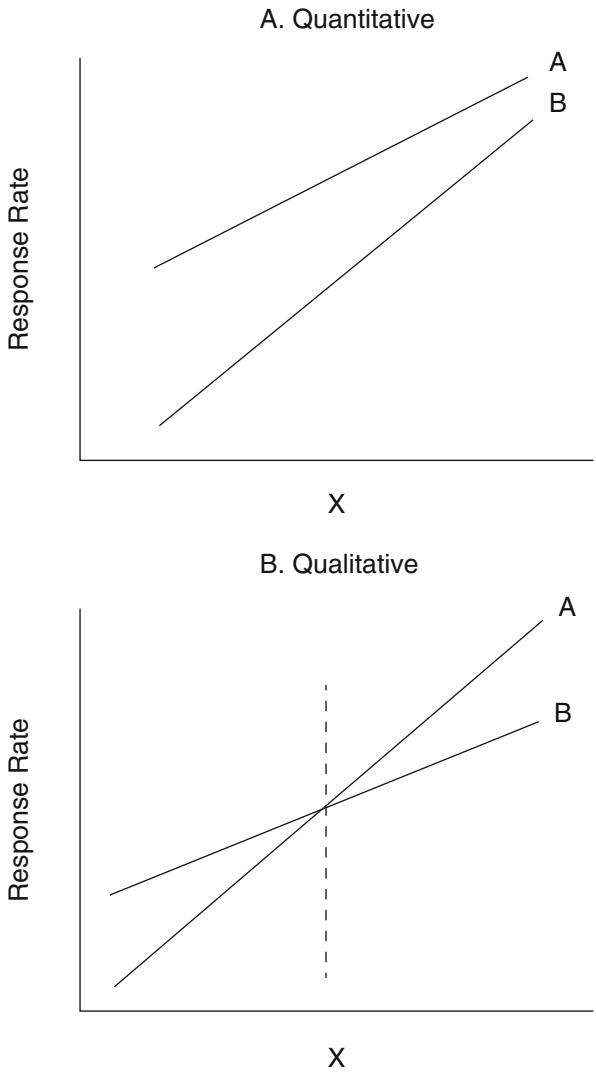

Fig. 18.5 Two types of intervention–covariate interactions [118]

changing direction, so a great deal of caution must be taken in the interpretation. One can test formally for interaction, but requiring a significant interaction test is much more cautious than reviewing the magnitude of intervention effect within each subgroup. Byar [105] presents a nice illustration example shown in Table 18.5. Two treatments, A and B, are being compared by the difference in mean response, $Y = \overline{X}_A - \overline{X}_B$, and S is the standard error of Y. In the upper panel, the interaction test is not significant, but examination of the subgroups is highly suggestive of interaction. The lower panel is more convincing for interaction, but we still need to examine each subgroup to understand what is going on.

Table 18.5 Examples of apparent treatment-covariate interactions [105]

Let $Y = \overline{X}_A - \overline{X}_B$			
	Statistic	*SE* of Y	P value (2 tail)
Unconvincing			
Overall test	$Y = 2S$	S	0.045
Subsets	$Y_1 = 3S$	$S\sqrt{2}$	0.034
	$Y_2 = 1S$	$S\sqrt{2}$	0.480
Interaction	$Y_1 - Y_2 = 2S$	$2S$	0.317
More convincing			
Overall test	$Y = 2S$	S	0.045
Subsets	$Y_1 = 4S$	$S\sqrt{2}$	0.005
	$Y_2 = 0$	$S\sqrt{2}$	1.000
Interaction	$Y_1 - Y_2 = 4S$	$2S$	0.045

Methods have been proposed for testing for overall interactions [114, 119]. However, Byar's concluding remarks [105] are noteworthy when he says,

> one should look for treatment-covariate interactions, but, because of the play of chance in multiple comparisons, one should look very cautiously in the spirit of exploratory data analysis rather than that of formal hypothesis testing. Although the newer statistical methods may help decide whether the data suffice to support a claim of qualitative interactions and permit a more precise determination of reasonable p values, it seems to me unlikely that these methods will ever be as reliable a guide to sensible interpretation of data as will medical plausibility and replication of the findings in other studies. We are often warned to specify the interactions we want to test in advance in order to minimize the multiple comparisons problem, but this is often impossible in practice and in any case would be of no help in evaluating unexpected new findings. The best advice remains to look for treatment-covariate interactions but to report them skeptically as hypotheses to be investigated in other studies.

As indicated in Chap. 6, the randomization in multicenter trials should be stratified by clinic. The analysis of such a study should, strictly speaking, incorporate the clinic as a stratification variable. Furthermore, the randomization should be blocked in order to achieve balance over time in the number of participants randomized to each group. These "blocks" are also strata and, ideally, should be included in the analysis as a covariate. However, there could be a large number of strata, since there may be many clinics and the blocking factor within any clinic is usually anywhere from four to eight participants. Use of these blocking covariates is probably not necessary in the analysis. Some efficiency will be lost for the sake of simplicity, but the sacrifice should be small.

As Fleiss [10] describes, clinics differ in their demography of participants and medical practice as well as adherence to all aspects of the protocol. These factors are likely to lead to variation in treatment response from clinic to clinic. In the Beta-blocker Heart Attack Trial (BHAT) [23], most, but not all, of the 30 clinics showed a trend for mortality benefit from propranolol. A few indicated a negative trend. In the Aspirin Myocardial Infarction Study (AMIS) [102], data from a few clinics suggested a mortality benefit from aspirin, although most clinics indicated little or no benefit. Most reported analyses probably do not stratify by clinic, but simply

combine the results of all clinics. However, at least one of the primary analyses should average within-clinic differences, an analysis that is always valid, even in the presence of clinic-treatment interaction [114].

Subgroup Analyses

While covariance or stratified analysis adjusts the overall comparison of main outcomes for baseline variables, another common analytic technique is to subdivide or subgroup the enrolled participants defined by baseline characteristics [128–156]. Here the investigator looks specifically at particular subgroups rather than the overall comparison to assess whether different groups of patients respond differently to the intervention. One of the most frequently asked questions during the design of a trial and when the results are analyzed is, "Are the intervention effects the same across levels of important baseline factors?" It is important that subgroups be examined. Clinical trials require considerable time and effort to conduct and the resulting data deserve maximum evaluation. Subgroup analyses can support or elaborate a trial's overall primary result, or provide exploratory results for the primary outcome that may have special interest for a particular subgroup. For example, analysis of data from the V-HeFT I trial suggested that the combination of isosorbide dinitrate and hydralazine might reduce mortality in blacks but not whites [157, 158]. This lead to the initiation of a follow-up trial of the combination which enrolled only blacks with advanced heart failure [159]. The A-HeFT trial concluded that this therapy increased survival [160]. However, such success stories are not common, and care must be exercised in the interpretation of subgroup findings.

How to perform subgroup analyses when reporting clinical trial data has long been a controversial topic [140, 156]. Manuscripts reporting the results of clinical trials commonly include statements about and estimates of effects in subgroups, but the results of subgroup analyses are often misleading, having been over-interpreted or presented in a way that makes their interpretation ambiguous [129, 149]. Most published advice since the early 1980's has included a common set of specific recommendations for subgroup analyses: they should be adjusted for multiple comparisons, they should be prespecified, and they should be assessed using interaction tests (rather than by within group estimates of the treatment effect) [142, 143, 153, 155, 156]. Making public a well-written protocol which specifies the proposed subgroups together with biologically plausible hypotheses for each and including plans for performing and presenting the subgroup analyses is often recommended as well.

As the number of subgroups increases, the potential for chance findings increases due to multiple comparisons [132, 143, 144, 155]. Therefore, if one were to perform tests of significance on a large number of subgroup analyses, there will be an increased probability of false positive results unless adjustments are made. Adjustment for multiple interaction tests on a set of variables defining

subgroups is necessary to control the number of false positive results. This can be done by such familiar methods as the Bonferroni correction or variants of it. An alternative suggested by guidelines for the New England Journal of Medicine is to report the expected number of false positives associated with a set of tests reported with nominal p-values [143, 153]; for example, this approach was taken for the ACCORD BP results [152]. Even with adjustments for multiplicity, however, over-interpretation of the results of treatment effects within subgroups can lead to irreproducible conclusions.

Ideally, the subgroups to be analyzed should be pre-specified. Since it is almost always possible to find at least one suggestive subgroup effect by persistent exploration of the data after a trial is over, even when the intervention is completely inert, defining the groups to be analyzed in advance, preferably with argument for their biological plausibility, confers the greatest credibility. There is likely to be, however, low power for detecting differences in subgroups [132, 155], and they are more likely to be affected by imbalances in baseline characteristics [161, 162]. Therefore, investigators should not pay as much attention to statistical significance for subgroup questions as they do for the primary question. Recognizing the low chance of seeing significant differences, descriptions of subgroup effects are often qualitative. On the other hand, as mentioned previously, testing multiple questions can increase the chance of a Type I error. Even when prespecified, there are reasons to be cautious.

The Clopidogrel for High Atherothrombotic Risk and Ischemic Stabilization, Management, and Avoidance (CHARISMA) trial [131] tested the effectiveness of long term dual antiplatelet therapy with clopidogrel plus low-dose aspirin to aspirin alone for the prevention of cardiovascular events among patients with either clinically evident CVD or multiple risk factors. Enrolled patients had either clinically evident cardiovascular disease (symptomatic) or multiple risk factors for atherothrombotic disease (asymptomatic). There was no difference between the two randomized arms, but 20 subgroup analyses were pre-specified in the protocol. For symptomatic vs. asymptomatic patients, the p-value for the interaction test was 0.045 and the p-value for benefit among the symptomatic patients was 0.046. This was reported as a suggestion of benefit for clopidogrel. Two accompanying editorials [143, 148] took issue with this conclusion for several reasons. The authors made no adjustment for multiple comparisons: had any correction been done, none of the subgroup analyses would have been even close to significant. The subsequent interpretation of the p-value for the symptomatic patients overstated its significance, which was marginal in any case. Furthermore the significance of the interaction test seemed to be driven more by a harmful effect in the asymptomatic patients than by any beneficial effect in the symptomatic patients. Finally, from the clinical point of view, the distinction between symptomatic and asymptomatic was not clear, since some of the patients in the asymptomatic group had a history of major cardiovascular events at baseline. A subsequent re-analysis of subgroups with patients identified as primary prevention and secondary prevention found no within-subgroup benefit for the primary endpoint [153].

Even if not explicitly pre-specified, subgroup analyses may be identified in several ways with different implications for the reliability of their results. For example, it might be reasonable to infer that subgroup hypotheses related to factors used to stratification of the randomization, such as age, sex or stage of disease, were in fact considered in advance. Factors that are integrated into the study design may be implied as subgroups even if they are not explicitly stated in the protocol.

Of course, the same problems in interpretation apply here as with formally prespecified subgroups. The Prospective Randomized Amlodipine Survival Evaluation Study (PRAISE), a large multicenter trial [146], pre-specified several subgroups, but in addition analyzed a baseline characteristic used to stratify the randomization, ischemic vs. non-ischemic etiology of chronic heart failure, as an additional subgroup. The randomization of participants with chronic heart failure was stratified by ischemic and non-ischemic etiology. While the primary outcome of death or cardiovascular hospitalization was nonsignificant and the secondary outcome of overall survival outcome was nearly significant ($p = 0.07$), almost all of the risk reduction was in the non-ischemic subgroup. The risk reduction was 31% for the primary outcome ($p = 0.04$) and 46% for mortality ($p < 0.001$). However, the more favorable result was expected to be in the ischemic subgroup, not the non-ischemic subgroup. Thus, the investigators recommended that a second trial be conducted in the patient population with non-ischemic chronic heart failure using a nearly identical protocol to confirm this impressive subgroup result [146]. The results of the PRAISE-II trial proved disappointing with no reduction in either the primary or secondary outcome [147]. Thus, the previous predefined subgroup result could not be confirmed.

On occasion, during the monitoring of a trial, particular subgroup findings may emerge and be of special interest. If additional participants remain to be enrolled into the trial, one approach is to test the new subgroup hypothesis in the later participants. With small numbers of participants, it is unlikely that significant differences will be noted. If, however, the same pattern emerges in the newly created subgroup, the hypothesis is considerably strengthened. Subgroups may also emerge during a trial by being identified by other, similar trials. If one study reports that the observed difference between intervention and control appears to be concentrated in a particular subgroup of participants, it is appropriate to see if the same findings occur in another trial of the same intervention, even though that subgroup was not pre-specified. Problems here include comparability of definition. It is unusual for different trials to have baseline information sufficiently similar to allow for characterization of identical subgroups. In the Raloxifene Use for The Heart (RUTH) [133], age groups were among a number of pre-specified subgroups, but the definition of the groups was modified to match what was used for the Women's Health Initiative [77]. Though the subgroup effects from RUTH and WHI were consistent, their interpretation as real clinical effects was vigorously challenged [155].

The weakest type of subgroup analysis involves post hoc analysis, sometimes referred to as "data-dredging" or "fishing." With this approach subgroups are suggested by the data themselves. Because many comparisons are theoretically

possible, tests of significance become difficult to interpret and should be challenged. Such analyses should serve primarily to generate hypotheses for evaluation in other trials. An example of subgrouping that was challenged comes from a study of diabetes in Iceland. Male children under the age of 14 and born in October were claimed to be at highest risk of ketosis-prone diabetes. Goudie [138] challenged whether the month of October emerged from post-study analyses biased by knowledge of the results. The ISIS-2 trial [141] illustrated a spurious subgroup finding that suggested treatment benefit of aspirin after myocardial infarction was not present in individuals born under Gemini or Libra astrological signs. A similar example [135] suggests twice as many participants with bronchial carcinoma were born in the month of March ($p < 0.01$) although this observation could not be reproduced [130, 134]. Subgroups unsupported by a biologically plausible hypothesis should be regarded with heightened caution.

Even subgroups supported by a biologically plausible rationale and suggesting beneficial effects can turn out to be irreproducible. Post-hoc subgroup analyses were performed for a number of trials of beta-blocking drugs were conducted in people who had had a myocardial infarction. One found that the observed benefit was restricted to participants with anterior infarctions [145]. Another claimed improvement only in participants 65 years or younger [128]. In the Beta-Blocker Heart Attack Trial, it was observed that the greatest relative benefit of the intervention was in participants with complications during the infarction [137]. These subgroup findings however, were not consistently confirmed in other trials [136].

Post-hoc subgroups may be specified by comparing participants from two groups who experience the same event, or have similar outcomes; an early example is the discriminant analysis done for the Multicentre International Study [145]. Investigators frequently want to do this in an attempt to understand the mechanisms of action of an intervention. Sometimes this retrospective look can suggest factors or variables by which the participants could be subgrouped. As discussed earlier in this chapter, categorization of participants by any outcome variable, e.g., adherence, can lead to biased conclusions. If some subgroup is suggested in this way, the investigator should create that subgroup in each randomized arm and make the appropriate comparison. For example, she may find that participants in the intervention arm who died were older than those in the control arm who died. This retrospective observation might suggest that age is a factor in the usefulness of the intervention. The appropriate way to test this hypothesis would be to subgroup all participants by age and compare intervention versus control for each age subgroup.

An interesting *post hoc* subgroup analyses was reported by the Metoprolol CR/XL Randomized Intervention Trial (MERIT) [154]. This trial, which evaluated the effect of a beta-blocker in participants with chronic heart failure, had two primary outcomes. One was all-cause mortality and the other was death plus hospitalization. MERIT was terminated early by the monitoring committee due to a highly significant reduction in mortality, as shown in Fig. 18.6, and similar reductions in death plus hospitalization. The results are remarkably consistent across all of the predefined subgroups for mortality, mortality plus hospitalization and mortality plus heart failure hospitalization as shown in Fig. 18.7. Moreover, the

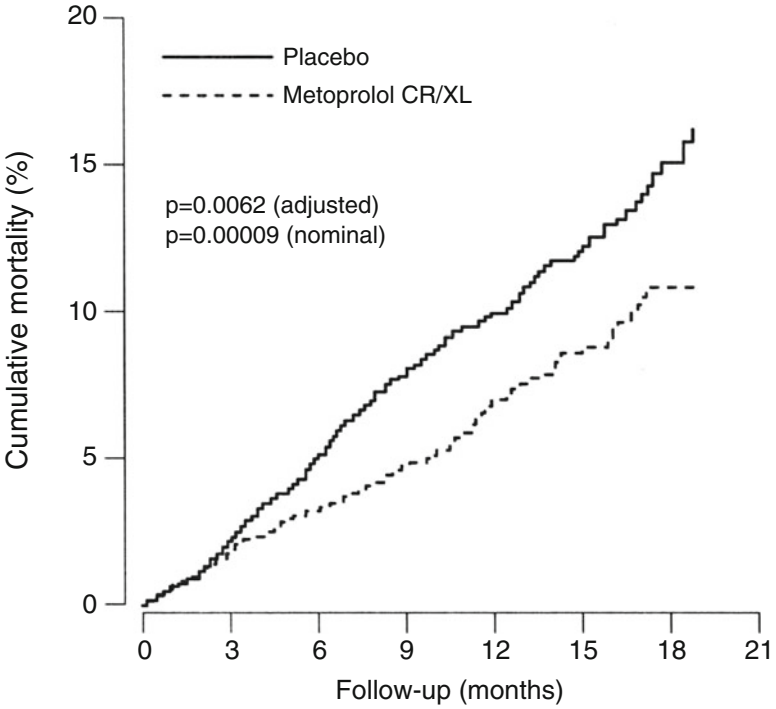

Fig. 18.6 MERIT Kaplan-Meier estimates of cumulative percentage of total mortality after randomization—p value nominal and adjusted for two interim analyses (MERIT) [37]. Reproduced with permission of Elsevier Ltd. for *Lancet*

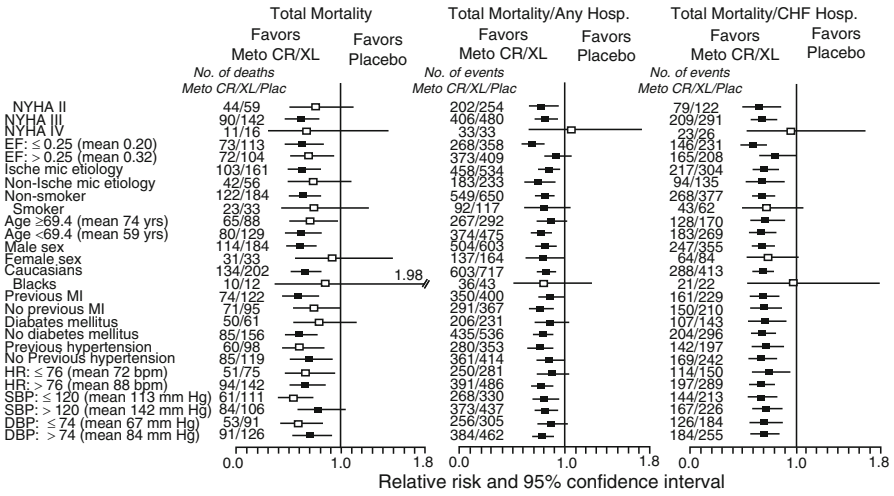

Fig. 18.7 Relative risk and 95% confidence intervals for selected subgroups in the MERIT trial, for total mortality, total mortality and all hospitalization, and total mortality and heart failure hospitalization [154]. Reproduced with the permission of Elsevier Ltd. for the *Amer Heart Journal*

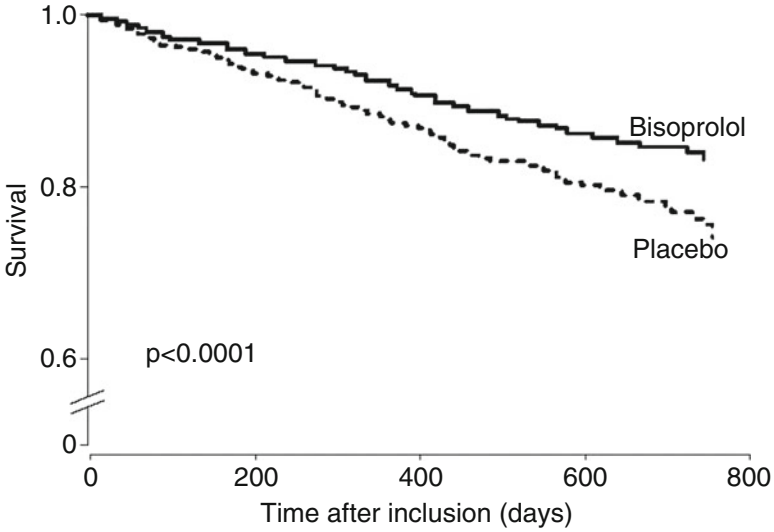

Fig. 18.8 Kaplan-Meier survival curves for the CIBIS-II trial, comparing bisoprolol and placebo [34]. Reproduced with the permission of Elsevier Ltd. for *Lancet*

results were very consistent with those from two other beta-blocker trials [37, 38], as shown in Figs. 18.8 and 18.9. However, post hoc analyses during review by regulatory agencies compared results among countries. These results are shown in Fig. 18.10. Of note is that for mortality, the relative risk in the United States slightly favors placebo, in contrast to the mortality results for the trial as a whole. With respect to outcomes of mortality plus hospitalization, and mortality and hospitalization for heart failure, the U.S. data are consistent with the overall trial results. As noted by Wedel et al. [154] the analysis for interaction depends on how the regional subgroups are formed. Whether the observed regional difference is due to chance or real has been debated, but Wedel and colleagues argued that is not consistent with other external data, not internally consistent within MERIT and not biologically plausible, and thus is most likely due to chance. This result does however point out the risks of post hoc subgroup analyses.

Regardless of how subgroups are selected, several factors can provide supporting evidence for the validity of the findings. As mentioned, similar results obtained in several studies strengthen interpretation. Internal consistency within a study is also a factor. If similar subgroup results are observed at most of the sites of a multicenter trial, they are more likely to be true. And of course, not all follow-up analyses and replication studies refute the initial subgroup finding. In contrast, however, plausible post hoc biological explanations for the findings, while necessary, are not sufficient. Given almost any outcome, reasonable sounding explanations can be put forward.

The two most common approaches to analysis of subgroup effects are (1) multiple hypothesis tests for effects within subgroups and (2) interaction tests for

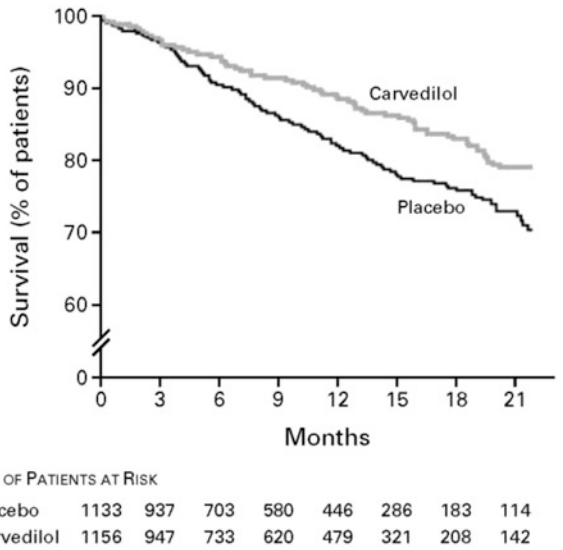

Fig. 18.9 Kaplan-Meier Analysis of Time to Death for COPERNICUS trial, comparing Placebo and Carvedilol Group. The 35% lower risk in the carvedilol group was significant: $p = 0.00013$ (unadjusted) and $p = 0.0014$ (adjusted) [38]

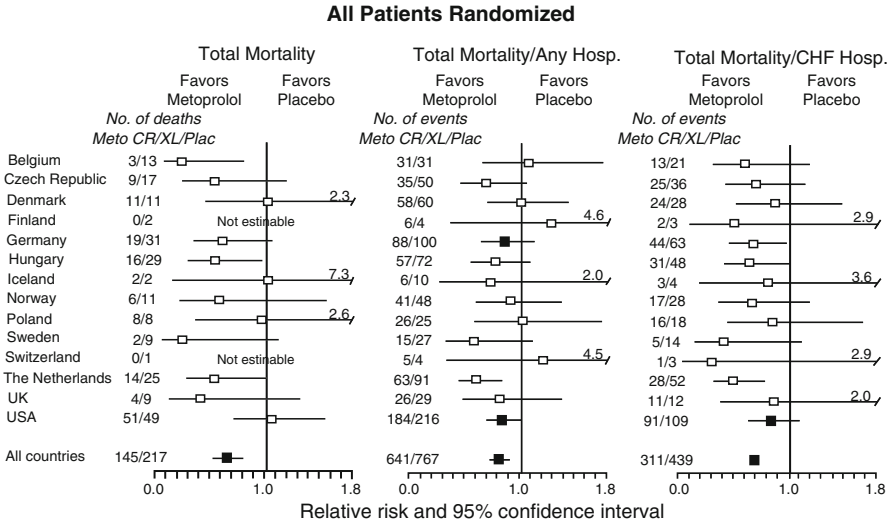

Fig. 18.10 Relative risk and 95% confidence intervals for the MERIT trial, for outcomes of total mortality, total mortality and hospitalization for any cause, and total mortality and heart failure hospitalization [154]. Reproduced with the permission of Elsevier Ltd. for the *Amer Heart Journal*

homogeneity of effects across subgroups defined by each variable of interest. Of these two the consensus in the literature strongly favors the interaction test. The interaction test provides a single, global assessment of whether a categorical variable partitioning the study cohort is associated with different magnitudes of

treatment effect. Estimates of those effects, with confidence intervals, provide exploratory indications of the consistency of the treatment effect across the population. Testing for treatment effects within subgroups, in contrast, requires a greater number of hypothesis tests and inflates the probability of a false positive result over the nominal significance level [132]. Statistical power and other considerations make the overall trial result a better guide to the effect in subgroups than the subgroup specific treatment effects [155, 156].

Often, attention is focused on subgroups with the largest intervention-control differences. However, even with only a few subgroups, the likelihood of large but spurious differences in effects of intervention between the most extreme subgroup outcomes can be considerable [136, 140, 150]. Because large, random differences can occur, subgroup findings may easily be over-interpreted. Peto has argued that observed quantitative differences in outcome between various subgroups are to be expected, and they do not imply that the effect of intervention is truly inhomogeneous [118].

It has also been suggested that, unless the main overall comparison for the trial is significant, investigators should be particularly conservative in reporting significant subgroup findings [150, 163]. Lee and colleagues conducted a simulated randomized trial, in which participants were randomly allocated to two groups, although no intervention was initiated [144]. Despite the expected lack of overall difference, a subgroup was found which showed a significant difference. Further simulations [132] have emphasized the potential for spurious results even when the main comparison is significant, and the importance of basing statements about significance on interaction tests rather than subgroup-specific tests.

In summary, subgroup analyses are important. However, they must be defined using baseline data and interpreted cautiously.

Not Counting Some Events

In prevention trials, the temptation is not to count events that are observed in the immediate post-randomization follow-up period. The rational for this practice is that events occurring that rapidly must have existed at screening, but were not detected. For example, if a cancer prevention trial randomized participants into a vitamin versus placebo trial, any immediate post randomization cancer events could not have been prevented since the cancer had to have already been present subclinically at entry. Because the intervention could not have prevented these cases, their inclusion in the design only dilutes the results and decreases power. While such an argument has some appeal, it must be viewed with caution. Rarely are mechanisms of action of therapies or interventions fully understood. More importantly, negative impact of interventions having a more immediate effect might not be seen as easily or as quickly with this approach. If used at all, and this should be rarely, the data must be presented both ways; i.e., with and without the excluded events.

An extreme case of dropping early events might be in a surgical or procedure trial. Participants assigned to the procedure might be put at higher risk of a fatal or irreversible event. These early risks to the participant are part of the overall intervention effect and should not be eliminated from the analysis.

Some trials have defined various counting rules for events once participants have dropped out of the study or reached some level of nonadherence. For example, the Anturane Reinfarction Trial [28] suggested that no events after 7 days going off study medication should be counted. It is not clear what length of time is appropriate to eliminate events to avoid bias. For example, if a participant with an acute disease continues to decline and is removed from therapy, bias could be introduced if the therapy itself is contributing to the decline due to adverse effects and toxicity. In the APPROVe trial [81–84] described earlier in this chapter, the decision not to count events after 14 days and not to follow participants after that period of time led to controversy. In fact, the results and the interpretation were different once the almost complete follow-up was obtained [84].

Comparison of Multiple Variables

If enough hypothesis tests are done, some may be significant by chance even if all the hypotheses being tested are false. This issue of multiple comparisons includes repeated looks at the same response variable (Chap. 15) and comparisons of multiple variables. Many clinical trials have more than one response variable, and prespecify several subgroups of interest. Thus, a number of statistical comparisons are likely to be made. For example, when performing 20 independent comparisons, one of them, on the average, will be significantly different by chance alone using 0.05 as the level of significance. The implication of this is that the investigator should be cautious in the interpretation of results if she is making multiple comparisons. The alternative is to require a more conservative significance level. As noted earlier, lowering the significance level will reduce the power of a trial. The issue of multiple comparisons has been discussed by Miller [164], who reviewed many proposed approaches.

One way to counter the problem is to increase the sample size so that a smaller significance level can be used while maintaining the power of the trial. However, in practice, most investigators could probably not afford to enroll the number of participants required to compensate for all the possible comparisons that might be made. As an approximation, if investigators are making k comparisons, each comparison should be made at the significance level α/k, a procedure known as the Bonferonni correction [164]. Thus, for $k = 10$ and $\alpha = 0.05$, each test would need to be significant at the 0.005 level. Sample size calculations involving a significance level of 0.005 will dramatically increase the required number of participants. The Bonferonni correction is quite conservative in controlling the

overall α level or false positive error rate if the test statistics are correlated, which is often the case. Therefore, it may be more reasonable to calculate sample size based on one primary response variable, limit the number of comparisons and be cautious in claiming significant results for other comparisons.

However, there are other procedures to control the overall α level and we summarize briefly two of them [165, 166]. Assume that we prespecify m hypotheses to be tested, involving multiple outcomes, multiple subgroups, or a combination. The goal is to control the overall α level. One implementation of the Holm procedure [166] is to order the p values from smallest to largest as $p(1)$, $p(2)$,. . ...,$p(m)$, corresponding to the m hypotheses $H(1)$, $H(2)$,. . .,$H(m)$. Then the Holm procedure would reject $H(1)$, if $p(1) \leq \alpha/m$. If and only if $H(1)$ is rejected can we consider the next hypothesis. In that case, $H(2)$ can be rejected if $p(2) \leq \alpha/(m-1)$. This process continues until we fail to reject and then the testing must stop. The Holm procedure can also be applied if the m hypotheses can be ordered according to their importance. Here, the most important hypothesis $H(1)$ can be rejected only if the corresponding p value is less than α/m. If rejected, the next most important hypothesis $H(2)$ can be rejected if the p value is less than $\alpha/(m-1)$.

Hochberg's procedure [165] also requires that the m hypotheses be specified in advance and orders the p-values from smallest to largest as does Holm's. The Hochberg procedure allows all m hypotheses to be rejected if $p(m) \leq \alpha/m$. If this is not the case, then the remaining $m-1$ hypotheses can be rejected if $p(m-1) \leq \alpha/(m-1)$. This process is carried out for all of the m hypotheses until a rejection is obtained and then stops. These two procedures will not give exactly the same rejection pattern so it is important to prespecify which one will be used.

In considering multiple outcomes or subgroups, it is important to evaluate the consistency of the results qualitatively, and not stretch formal statistical analysis too far. Most formal comparisons should be stated in advance. Beyond that, one engages in observational data analysis to generate ideas for subsequent testing.

Use of Cutpoints

Splitting continuous variables into two categories, for example by using an arbitrary "cutpoint," is often done in data analysis. This can be misleading, especially if the cutpoint is suggested by the data. As an example, consider the constructed data set in Table 18.6. Heart rate, in beats per minute, was measured prior to intervention in two groups of 25 participants each. After therapies A and B were administered, the heart rate was again measured. The average changes between groups A and B are not sufficiently different from each other ($p = 0.75$) using a standard t-test. However, if these same data are analyzed by splitting the participants into "responders" and "non-responders," according to the magnitude of heart rate reduction, the results can be made to vary. Table 18.7 shows three such possibilities, using

Table 18.6 Differences in pre- and post-therapy heart rate, in beats per minute (HR), for Groups A and B, with 25 participants each

	A			B		
Observation number	Pre HR	Post HR	Change in HR	Pre HR	Post HR	Change in HR
1	72	72	0	72	70	2
2	74	73	1	71	68	−3
3	77	71	6	75	74	1
4	73	78	−5	74	71	3
5	70	66	4	71	73	−2
6	72	76	−4	73	78	−5
7	72	72	0	71	69	2
8	78	76	2	70	74	−4
9	72	80	−8	79	78	1
10	78	71	7	71	72	−1
11	76	70	6	78	79	−1
12	73	77	−4	72	75	−3
13	77	75	2	73	72	1
14	73	79	−6	72	69	3
15	76	76	0	77	74	3
16	74	76	−2	79	75	4
17	71	69	2	77	75	2
18	72	71	1	75	75	0
19	68	72	−4	71	70	1
20	78	75	3	78	74	4
21	76	76	0	75	80	−5
22	70	63	7	71	72	−1
23	76	70	6	77	77	0
24	78	73	5	79	76	3
25	73	73	0	79	79	0
Mean	73.96	73.20	.76	74.40	73.96	0.44
Standard deviation	2.88	3.96	4.24	3.18	3.38	2.66

Table 18.7 Comparison of change in heart rate in Group *A* versus *B* by three choices of cutpoints

Beats/min	<7	≥7	<5	≥5	<3	≥3
Group *A*	25	2	19	6	17	8
Group *B*	25	0	25	0	18	7
Chi-square	$p = 0.15$		$p = 0.009$		$p = 0.76$	
Fisher's exact	$p = 0.49$		$p = 0.022$		$p = 0.99$	

reductions of 7, 5, and 3 beats per minute as definitions of response. As indicated, the significance levels, using a chi-square test or Fisher's exact test, change from not significant to significant and back to not significant. This created example suggests that by manipulating the cutpoint one can observe a significance level less than 0.05 when there does not really seem to be a difference.

Noninferiority Trial Analysis

As discussed in Chap. 5, noninferiority trials are challenging to design, conduct and analyze. We pointed out the special challenges in setting the margin of noninferiority. However, once that margin of noninferiority is established prior to the start of the trial, there remain several issues that must be included in a rigorous analysis and reported because of the clinical and regulatory implications [13, 167–183]. If we define I to be the new intervention, C to be the control or standard, and P to be the placebo or no treatment, then we obtain from the noninferiority trial an estimate of the relative risk (RR) of I to C, RR(I/C) or an absolute difference. In the design, the metric must be established since the sample size and the interim monitoring depend on it. The first analytic challenge is to establish whether the new intervention met the criteria for noninferiority, a part of which is demonstrating that the upper limit of the 95% confidence interval of the estimate was less than the noninferiority margin.

As shown in Fig. 18.11, from Pocock and Ware [181], if the upper limit of the 95% confidence interval for the relative risk is less than unity, various degrees of evidence exist for superiority (See case A). For noninferiority trials, if the upper limit of the 95% confidence interval is less than the margin of non-inferiority, δ, then there is evidence for noninferiority (see cases B and C). Failure to be less than this margin does not provide evidence for noninferiority (see case D). The design must have sufficient sample size and power to rule out a margin of noninferiority as discussed in Chap. 8. Although not expected when the study was designed, a noninferiority trial might also indicate harm (See E).

The second desired goal of a noninferiority analysis is to demonstrate that the new intervention would have beaten a placebo or no treatment if it had been included; that is, the estimate of RR(I/P). Analytically, this can be accomplished by recognizing that RR(I/P) = RR(I/C) RR(C/P). However, for this imputation step to work requires at least two critical assumptions: (1) there is constancy of the control effect over time, and (2) the population where the control was tested against placebo is relevant to the current use where the intervention (I) is being tested. These assumptions are difficult, perhaps impossible, to establish (see Chap. 5). In this chapter, we will focus our attention on the first challenge of establishing whether or not the intervention versus control comparison was less than the noninferiority margin.

Assuming that an appropriate active control was selected, the trial must implement it according to best practice and as good or better than that what was done in the initial trial establishing its benefit [172]. Otherwise, the new intervention is being compared to a control that is handicapped, making it easier for the new intervention to appear similar or even better than the control. Poor adherence and conduct will favor the new intervention in a noninferiority trial, instead of handicapping the new intervention as in a superiority trial [179]. Thus, as discussed in Chap. 14, adequate measures of adherence must be collected during the trial in order to make this critical assessment. Adherence in this case does not only mean

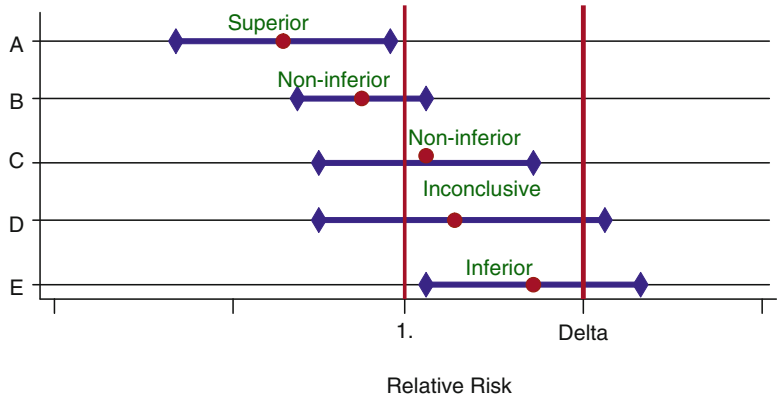

Fig. 18.11 Relative risks and 95% confidence intervals for a series of superiority and non-inferiority trials [181]

whether the participant took all or almost all of the intervention and control drugs. What else participants were taking as concomitant medication is also a consideration. If there is a substantial imbalance, interpretation of the results would be difficult.

Another key factor is whether the outcomes chosen are true measures of the effect of both the new intervention and the control. This is sometimes referred to as *assay sensitivity* [177]. Thus, whether consciously or not, an investigator might select an outcome that would show no change no matter what intervention was being studied, and thus guarantee that the noninferiority margin would be achieved. Outcomes should be similar to those used in the positive control versus placebo trials.

There is a debate whether the intention-to-treat analysis or the "on treatment" analysis is most appropriate for a noninferiority designed trial. If intention-to-treat is used, nonadherence dilutes whatever differences may exist and thus is biased towards noninferiority. An "on treatment" analysis compares only those who are good adherers, or at least took some predefined portion of the intervention and thus is closer to testing the true effect. However, as we demonstrated earlier in this chapter, analyzing trials by adherence to an intervention can be substantially biased, the direction of which cannot be predicted. Thus, we do not recommend such an analysis because of the uncertainty of bias and its direction, and instead recommend that a trial be designed to minimize nonadherence. The true comparison of the new intervention may be somewhere in between the intention-to-treat and the "on treatment" but there is no dependable way to tease that estimate out. If both analytic approaches confirm noninferiority, then the conclusion is more robust, assuming that the noninferiority margin is reasonable [178].

Any trial relies on an adequate sample size to have power to test hypotheses of interest, whether for superiority or noninferiority. For a superiority trial, inadequate sample size works against finding differences but for noninferiority, inadequate sample size favors finding noninferiority. There is a difficult balance between

having a noninferiority margin that is too small and thus requiring an unachievable sample size and having a margin that is so large that the sample size is appealing but the results would not be convincing.

There are many examples of noninferiority trials but we will use one to illustrate the challenges. The Stroke Prevention using an ORal Thrombin Inhibitor in atrial Fibrillation (SPORTIF)-V trial in participants with atrial fibrillation comparing a new intervention, ximelegatran, against a standard warfarin intervention [183], with a primary outcome of stroke incidence. A number of issues were involved. First, there were no very good warfarin versus placebo trials to set the noninferiority margin. Second, the trial used absolute difference as the metric, assuming the event rate would be around 3% , but instead observed an event rate less than half that. Thus, the noninferiority margin of 2% that was prespecified was too large given the small event rate. If the observed event rate of 1.5% had been assumed, the prespecified margin would have been much less, perhaps closer to 1%. The observed stroke rates were 1.2% in the warfarin group and 1.6% in the ximelegatran group with a 95% CI of −.13% to 1.03% which would meet the initial margin of noninferiority. However, this was not adequate for a margin of 1%. Therefore, even though margins may be set in advance, results may invalidate the assumptions and thus the margin itself.

As discussed in Chap. 21, presentation of the results of noninferiority trials are more complex than for superiority trials because all of the assumptions must be so carefully and clearly laid out [181].

Analysis Following Trend Adaptive Designs

As discussed in Chaps. 5 and 17, the design of a trial may have an adaptive element. This might be a group sequential design for early termination due to overwhelming benefit or a strong signal for harm, or perhaps futility. Among the adaptive designs discussed some involved changing the sample size. Some of these sample size changes are due to overall lower event rates or higher variability in the primary outcome than was assumed in the original sample size estimate. In these instances, the final analysis proceeds as normal. However, another method for sample size change relies on trend adaptive designs. In these designs, which depend on the emerging trend in the data, the final critical value or significance level will be affected and thus must be kept in mind for the final analysis.

For example, some trials may monitor accumulating interim data and may terminate the trial early for evidence of benefit or harm. If a group sequential design using a 0.05 two-sided significance level O'Brien-Fleming boundary were used five times during the trial, approximately equally spaced, the final critical value would not be +1.96 and −1.96 for the upper and lower bounds but a value closer to 2.04.

For trend adaptive sample size changes, the final critical value depends on which methodology was used but all will require typically a more conservative value, for example, than a two-sided nominal alpha level of 0.05 (a critical value of 1.96).

Other than adjusting the final critical value, the analyses for these trend adaptive designs may also utilize a modified test statistic. For example, if the method of Cui et al. [184] is used in increasing the sample size, a weighted test statistic as described in Chap. 17 is required. Future observations are given less weight than the early existing observations. The usual test statistic is not appropriate in this situation. For the other trend adaptive methods described in Chaps. 5 and 17, the final analysis can proceed with the standard statistics in a usual straightforward fashion, adjusting for the final critical value from sequential testing as appropriate.

Meta-analysis of Multiple Studies

Often in an area of clinical research several independent trials using similar participants and similar intervention strategies are conducted over a period of a few years. Some may be larger multicenter trials, but there may be a substantial number of small trials none of which were conclusive individually, though they may have served as a pilot for a larger subsequent study. Investigators from a variety of medical disciplines often review the cumulative data on similar trials and try to develop a consensus conclusion of the overall results [185–193]. If this overview is performed by a formal process and with statistical methods for combining all the data with a single analysis, the analysis is usually referred to as a meta-analysis or systematic review. Methods suitable for this purpose were described in 1954 by Cochran [194] and later by Mantel and Haenszel [195]. Other authors have summarized the methodologic approaches [196–207]. The Cochrane Collaboration has been a major contributor to systematic reviews of controlled trials [208], often organized around a specific health care area or issue, including systematic reviews of adverse effects and advice on how to conduct such systematic reviews. Guidelines intended to improve the conduct and reporting of meta-analyses have been published [209, 210]. There are numerous examples of meta-analysis in a variety of medical disciplines and a few are referenced here [211–221]. A great deal has been written and discussed about the usefulness and challenges of meta-analyses [222–233].

Rationale and Issues

Researchers conduct systematic reviews and meta-analyses to address a number of important questions [190]. Probably the most common reason is to obtain more precise estimates of an intervention effect and to increase the power to observe small but clinically important effects. Very often the potential for increased power

to detect small but clinically important effects motivates the meta-analysis. However, meta-analyses can also evaluate the generalizability of results across trials, populations, and specific interventions. Subgroup analyses based on small numbers of participants may not lead to firm conclusions and miss qualitative differences in effect. Post hoc subgroup analyses are unreliable due to multiplicity of testing. Prespecified meta-analysis offers the opportunity to examine a limited number of hypotheses identified in individual trials. Meta-analysis of subgroups can guide clinicians in their practice by selecting participants most suitable for the intervention. In addition, meta-analysis can support submissions to the U.S. Food and Drug Administration. If a major clinical trial is being initiated, a sensible approach is to base many aspects of the design on the summary of all existing data. Meta-analysis is a systematic process that can provide critical information on definitions of population and intervention, control group response rates, expected size of the intervention effect, and length of follow-up. Finally, if a new treatment or intervention gains widespread popularity early in its use, a meta-analysis may provide a balanced perspective and may suggest the need for a single, large, properly designed clinical trial to provide a definitive test. Furthermore, meta-analyses are mandated if the opportunity to conduct a new large study no longer exists due to a loss of equipoise, even if this loss is not well justified. In this case, a meta-analysis may be the only solution for salvaging a reliable consensus.

As indicated, a meta-analysis is the combination of results from similar participants evaluated by similar protocols and interventions. The standard analysis of a multicenter trial, stratified by clinical center is in some ways an ideal meta-analysis. Each center plays the role of a small study. Protocols and treatment strategies are identical, and participants are more similar than those in a typical collection of trials.

This contrast between a meta-analysis and a multicenter trial points out some limitations of the former. While the implementation of a clinical protocol can vary across centers, such differences are negligible compared to those in a collection of independently conducted large or small trials. Even when the analysis is done by pooling participant-level data from each trial [212, 217], meta-analysis cannot be expected to produce the same level of evidence as a single, large clinical trial. In a typical meta-analysis, important differences exist in actual treatment, study population, length of follow-up, measures of outcome, level of background medical care in international trials and quality of data [222, 225–228, 233]. Because of these differences, the potential for meta-analysis should never be a justification for conducting a series of small, loosely connected studies with the expectation that a definitive result can be produced by combining after the fact. Perhaps the most fundamental problem is the potential to create bias when deciding on which studies to include in a meta-analysis. Two examples of such bias are selection bias and investigator bias.

Many support the concept that the most valid overview and meta-analysis requires all relevant studies conducted be available for inclusion or at least for consideration [190, 226]. Failing to do so can produce selection bias; that is, a mis-estimation caused by analysis of a non-representative sample. For example,

Furberg [228] provides a review of seven meta-analyses of lipid lowering trials. Each article presents different inclusion criteria, such as the number of participants or the degree of cholesterol reduction. The results vary depending on the criteria used. Another example of selection bias in meta-analysis involves the investigation of whether adding manual thrombus aspiration to primary percutaneous coronary intervention (PPCI) reduces total mortality. Between 1996 and 2009, about 20 small clinical trials and one larger trial, the Thrombus Aspiration during Percutaneous Coronary Intervention in Acute Myocardial Infarction Study (TAPAS) trial [234], were conducted to address whether PPCI with thrombus aspiration might have benefits over PPCI alone. These trials were not powered for total mortality and the smaller trials were not consistently positive; however, the largest suggested a possible 50% mortality benefit for manual thrombus aspiration. A series of meta-analyses sought to clarify the situation [212, 235–240]. Despite having identical aims, nearly identical inclusion criteria, and access to the same small set of trial results, no two meta-analyses included the same set of studies, and results varied. Because there were conflicting conclusions, no consensus was produced. The Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (TASTE) trial, designed with mortality as its primary outcome, concluded that there was no effect [20, 241], but a subsequent meta-analysis including the TASTE trial, while finding a non-significant effect on mortality, concluded that a modest reduction in clinical outcomes exists [242].

While it is clearly difficult enough to decide which well-known and published trial results to include, a further serious complication is that some relevant trial results may not be readily accessible in the literature due to publication bias [223, 231]. Published trials are more likely to be statistically significant ($p < 0.05$) or to favor a novel intervention. Trials that yield neutral or indifferent results are less likely to be published. One example described by Furberg and Morgan [227] illustrates this problem. An overview [223] of the use of propranolol in patients following a heart attack reported 7 of 45 patients died in the hospital compared to a non-randomized, placebo-control where 17 of 46 died, indicating a clear benefit of propranolol. Controversy over design limitations motivated the investigator to conduct two additional randomized trials. One showed no difference and the other a negative (harmful) trend. Neither was ever published. Identifying yet another obstacle to inclusion of all relevant studies, Chalmers et al. [224] pointed out that a MEDLINE literature search may only find 30–60% of published trials. This is due in part to the way results are presented and searches of typical key words may not uncover relevant papers. Although search engines may be better now, there are undoubtedly still limitations. Work by Gordon and colleagues found that only 57% of 244 NHLBI-supported trials completed between January 2000 and December 2011 published their main results within 30 months after completion [243]. These difficulties in determining and accessing the entire population of relevant studies may lead to analysis of a subset of trial results which are not representative, producing conclusions which do not reflect the totality of evidence because of selection bias.

Another type of bias, referred to as investigator bias, occurs when an investigator ignores or goes beyond any pre-specified plan and makes subjective decisions about which trials and outcome variables will get reported. If protocols were written well and adhered to strictly, investigator bias would not be a problem. However, post-hoc repeated testing of multiple subgroups and multiple outcomes may not be easy to detect from the published report [229]. Promising early results may draw major attention, but if later results show smaller intervention effects, they may go unnoticed or be harder to find for the systematic review. Furthermore, authors of systematic reviews are also to subject to investigator bias. That is, unless the goals of the meta-analysis are clearly stated a priori in a protocol, a positive result can be found in this analysis by sifting through numerous attempts. A great deal of time and persistence are required in order to get access to all known conducted trials and accurately extract the relevant data. Not all meta-analyses are conducted with the same degree of thoroughness.

The medical literature is filled with meta-analysis of trials covering a wide range of disciplines [211–221]. Several examples from the cardiology literature will provide an overview. Chalmers and colleagues [214] reviewed six small studies that used anticoagulants in an effort to reduce mortality in heart attack patients. While only one of the six was individually significant, the combined overall results suggested a statistically significant 4.2% absolute reduction in mortality. The authors suggested no further trials were necessary. However, due to issues raised, this analysis drew serious criticism [229]. Several years later, Yusuf and colleagues [221] reviewed 33 fibrinolytic trials, focusing largely on the use of streptokinase. This overview included trials with much dissimilarity in dose, route and time of administration, and setting. Although the meta-analysis for intravenous use of fibrinolytic drugs was impressive, and the authors concluded that results were not due to reporting biases, they nevertheless discussed the need for future large-scale trials before widespread use should be recommended. There were issues, for example, as to how quickly such an intervention needed to be started after onset of a heart attack. That is, timing needed to be resolved. Canner [213] conducted an overview of 6 randomized clinical trials testing aspirin use in participants with a previous heart attack to reduce mortality. His overall meta-analysis suggested a 10% reduction that was not significant ($p = 0.11$). However, there was an apparent heterogeneity of results and the largest trial had a slightly negative mortality result. The Canner overview was repeated by Hennekens et al. [215] after several more trials had been conducted. This updated analysis demonstrated favorable results. May et al. [218] conducted an early overview of several modes of therapy for secondary prevention of mortality after a heart attack. Their overview covered anti-arrhythmic drugs, lipid-lowering drugs, anticoagulant drugs, beta-blocker drugs, and physical exercise. Although statistical methods were available to combine studies within each treatment class, they chose not to combine results, but simply provided relative risks and confidence interval results graphically for each study. A visual inspection of the trends and variation in trial results suggests a summary analysis. Yusuf et al. [220] later provided a more detailed overview of beta blocker trials. While using a similar graphical presentation, they calculated a summary odds

ratio and its confidence interval. Meta-analysis of cancer trials have also been conducted including the use of adjuvant therapy for breast cancer [216]. While using multiple chemotherapeutic agents indicated improved relapse-free survival after 3 and 5 years of follow-up, as well as for survival, the dissimilarity among the trials led the authors to call for more trials and better data.

Thompson [232] pointed out the need to investigate sources of heterogeneity. These differences may be in populations studied, intervention strategies, outcomes measured, or other logistical aspects. Given such differences, inconsistent results among individual studies might be expected. Statistical tests for heterogeneity often have low statistical power even in the presence of a moderate heterogeneity. Thompson [232] argued that we should investigate the influence of apparent clinical differences between studies and not rely on formal statistical tests to give us assurance of no heterogeneity. In the presence of apparent heterogeneity, overall summary results should be interpreted cautiously. Thompson described an example of a meta-analysis of 28 studies evaluating cholesterol lowering and the impact on risk of coronary heart disease. A great deal of heterogeneity was present, so a simple overall estimate of risk reduction may be misleading. He showed that factors such as age of the cohort, length of treatment, and size of study were contributing factors. Taking these factors into account made the heterogeneity less extreme and results more interpretable. One analysis showed that the percent reduction in risk decreased with the age of the participant at the time of the event, a point not seen in the overall meta-analysis. However, he also cautioned that such analyses of heterogeneity must be interpreted cautiously, just as for subgroup analyses in any single trial.

Meta-analysis, as opposed to typical literature reviews, usually puts a p-value on the conclusion. The statistical procedure may allow for calculation of a p-value, but it implies a precision which may be inappropriate. The possibility that not all relevant studies have been included may make the interpretation of the p-value tenuous. Quality of data may vary from study to study. Data from some trials may be incomplete without being recognized as such. Thus, only very simple and unambiguous outcome variables, such as all-cause mortality and major morbid events ought to be used for meta-analysis.

Statistical Methods

Since meta-analysis became a popular approach to summarizing a collection of studies, numerous statistical publications have been produced addressing technical aspects [186, 194–196, 198–201, 203, 205, 207]. Most of this is beyond the technical scope of this text, but a number of texts on the subject of meta-analysis are available [197, 202, 204, 206]. Two common technical approaches were first suggested by Cochran [194] in 1954. If all trials included in the meta-analysis are estimating the same true (but unknown) fixed effect of an intervention, the Mantel-Haenszel method [195] can be used with a slight variation. This is similar to the

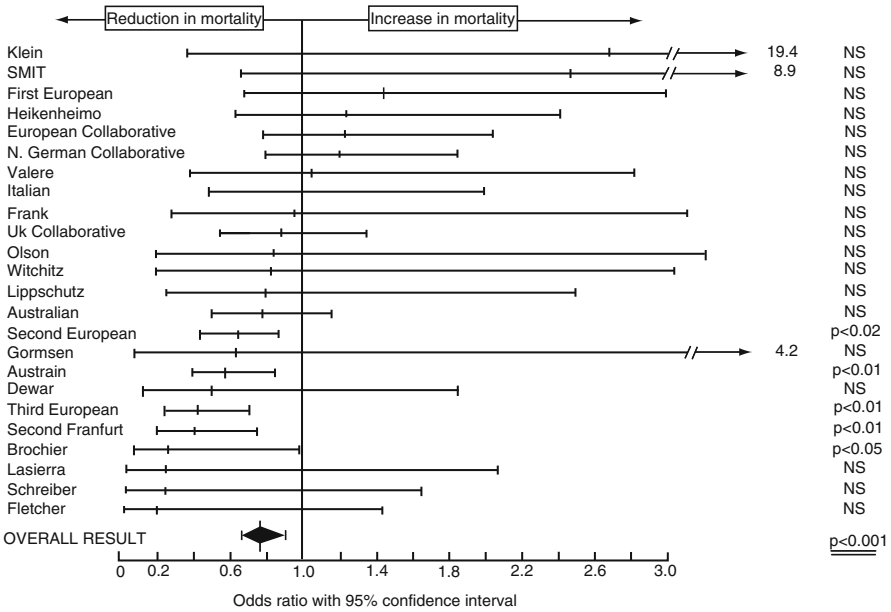

Fig. 18.12 Apparent effects of fibrinolytic treatment on mortality in the randomized trials of IV treatment of acute myocardial infarction (Reproduced with permission of the Editor, European Heart Journal and Dr. S. Yusuf)

logrank or Mantel-Haenszel method in the chapter on survival analysis. If the trials are assumed to have dissimilar or heterogeneous true intervention effects, the effects are described by a random effects model, as suggested by DerSimonian and Laird [200]. Another valid but less common approach relies on a Bayesian analysis [204] which was used to assess the literature on adjunctive thrombotomy for acute myocardial infarction [239].

The method of DerSimonian and Laird [200] compares rate differences within each study, and obtains a pooled estimate of the rate difference as well as the standard error. The pooled estimate of the rate difference is a weighted average of the individual study rate differences. The weights are the inverse of the sum of the between and within study variance components of intervention effect. If the studies are relatively similar or homogeneous in intervention effect, this approach and the fixed effects method produce very similar results [196]. Heterogeneity tests generally are not as powerful as the test for main effects. However, if studies vary in intervention effect, these two methods can produce different results as illustrated by Berlin et al. [196] as well as Pocock and Hughes [203].

Typically, when presenting the results of a meta-analysis, the OR estimate and 95% confidence interval are plotted in a single graph for each trial to provide a visual summary. Figure 18.12, from Yusuf et al. [221], summarizes the effects of 24 trials of fibrinolytic treatment on mortality in people with an acute heart attack.

The hash mark represents the estimated OR and the line represents the 95% confidence interval. They [221] include a single estimate of the OR, combining all studies. The size of the symbol in this plots, sometimes referred to as "forest plots," is an indication of the size of each individual studies. In the presence of serious heterogeneity of treatment effect, however, the appropriateness of obtaining a single point estimate must be questioned. If the heterogeneity is qualitative; that is, some estimates of the OR are larger than unity and others less than unity, then a combined single estimate is perhaps not wise. This would be especially true if these estimates indicated a time trend, which could occur if dose and participant selection changed as more experience with the new intervention was obtained.

Which model to use for meta-analysis is a matter of debate, but none are exactly correct. The random effects model has an undesirable aspect, in that small trials may dominate the final estimate. With the fixed effect model, larger trials get greater weight. Since the meta-analysis is conducted on available trials, however, the sample of participants included is not likely to be very representative of the general population to which the intervention may be applied. That is, the trials that are available do not contain a random sample of people from the targeted population but rather are participants who volunteered and who in other respects may not be representative. Thus, the estimate of the intervention effect is not as relevant as whether or not the intervention has an effect. We prefer a fixed effects model but suggest that both models should be conducted to examine what, if any, differences exist.

Chalmers, a strong advocate of clinical trials, argued that participants should be randomized early in the evolution and evaluation of a new intervention [244]. Both as a result of that kind of advocacy and the fact that small trials are always done before large ones in the development of new interventions, an early meta-analysis is likely to consist of many small studies. Sometimes, meta-analyses of just small trials might yield significant results.

Thus, meta-analyses are seen by many as alternatives to the extraordinary effort and cost often required to conduct adequately powered individual trials. Rather than providing a solution, they perhaps ought to be viewed as a way of summarizing existing data; a way that has strengths and weaknesses, and must be critically evaluated. It would clearly be preferable to combine resources prospectively and collaborate in a single large study. Pooled results from distinct studies cannot replace individual, well-conducted multicenter trials.

Analysis for Harmful Effects

While the analyzing the primary and secondary outcome variables for benefit is challenging, the analysis of adverse event data for safety is even more complex and challenging. Of course, if any of the primary or secondary outcome variables trend in the wrong direction, then there is evidence of harm, not benefit. However, harmful effects may manifest themselves in other variables than these primary or

secondary outcomes. Some adverse event measures can be prespecified such as changes in the QT interval in an ECG or an elevated liver function test (LFT). But there are many other possibilities.

The typical way that adverse event data are collected in current Phase III trials is a passive system where patient complaints or physician observations are summarized in text fields which are later coded by various adverse event coding systems (See Chap. 12). Such events are usually not solicited actively so that if the patient does not complain or the physician does not record the event or problems, they do not get coded. In fact, if a patient complains about the adverse event in a different manner from one visit to the next, the event may be coded differently. If the physician records the event using different language, the event may get coded differently. It can be challenging to even track an adverse event from one visit to the next within a patient. Another one of the problems of these types of coding systems is that a very large number of categories can be generated for the same essential problem, depending on how the patient complained or the physician recorded his observations in the patient chart.

Thus, tables of adverse events using these systems can have very many rows with only a few events in each row, even for the same basic adverse problem. Such data are not likely to produce statistically significant comparisons or flag potential problems. The data are so granular that an adverse event signal cannot be seen easily. These coding systems can collapse these detailed categories into higher order terms but in doing so add adverse events that are a real signal with typically a lot more events that are not very serious or clinical important. That is, the noise drowns out the signal.

Thus, analysis of this type of data requires a careful scrutiny of the numerous detailed categories to find ones that seem to indicate a meaningful clinical issue, and these items may come from different higher level categories. This process is or can be very subjective and may be hard for another investigative team to reproduce this same categorization.

One alternative to this passive adverse event reporting is to specify in the protocol the special adverse events of interest, and actively solicit the participants for information on their occurrence or conduct whatever laboratory measures are necessary to assess whether that event did occur. Examples of a deal breaker might be QT interval increase or an increase in LFT measures. Any substantial, statistically significant or clinically important imbalance in these type of events would be sufficient to perhaps terminate a trial early or kill the further development of the intervention, whether drug, device or biologic. There are probably more than 10 such “deal breakers” but less than 100, depending on the disease and intervention. Of course other adverse event data may be collected in a patient chart as text and later retrieved as necessary using more recent developed natural language processing (NLP) algorithms. If imbalances are found in such review, confirmation should be sought whenever possible using warehouse data from large electronic health record (EHR) systems.

References

1. Geller NL. Advances in Clinical Trial Biostatistics. Taylor & Francis, 2003.
2. Van Belle G, Fisher LD, Heagerty PJ, Lumley T. Biostatistics: A Methodology For the Health Sciences. Wiley, 2004.
3. Piantadosi S. Clinical Trials: A Methodologic Perspective. Wiley, 2013.
4. Cook TD, DeMets DL. Introduction to Statistical Methods for Clinical Trials. Taylor & Francis, 2007.
5. Pagano M, Gauvreau K. Principles of Biostatistics. Duxbury, 2000.
6. Hill AB. Principles of medical statistics. Oxford University Press, 1971.
7. Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research. Wiley, 2008.
8. Woolson RF, Clarke WR. Statistical Methods for the Analysis of Biomedical Data. Wiley, 2011.
9. Armitage P. The analysis of data from clinical trials. *Statistician* 1979;171–183.
10. Fleiss JL. Analysis of data from multiclinic trials. *Control Clin Trials* 1986;7:267–275.
11. Newcombe RG. Explanatory and pragmatic estimates of the treatment effect when deviations from allocated treatment occur. *Statist Med* 1988;7:1179–1186.
12. 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.
13. Temple R, Ellenberg SS. Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments. Part 1: Ethical and Scientific Issues. *Ann Intern Med* 2000;133:455–463.
14. Jones B, Jarvis P, Lewis JA, Ebbutt AF. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996;313:36–39.
15. Sackett DL, Gent M. Controversy in Counting and Attributing Events in Clinical Trials. *N Engl J Med* 1979;301:1410–1412.
16. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. *J Chronic Dis*1967;20:637–648.
17. Gillespie SH, Crook AM, McHugh TD, et al. Four-Month Moxifloxacin-Based Regimens for Drug-Sensitive Tuberculosis. *N Engl J Med* 2014.
18. FDA: International Conference on Harmonization - Efficacy: Statistical principles for clinical trials. U S Food and Drug Administration.
19. Lewis JA. Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline. *Statist Med* 1999;18:1903–1942.
20. Frobert O, Lagerqvist B, Olivecrona GrK, et al. Thrombus Aspiration during ST-Segment Elevation Myocardial Infarction. *N Engl J Med* 2013;369:1587–1597.
21. The Women's Health Initiative Study Group. Design of the Women's Health Initiative Clinical Trial and Observational Study. *Control Clin Trials* 1998;19:61–109.
22. May GS, DeMets DL, Friedman LM, et al. The randomized clinical trial: bias in analysis. *Circulation* 1981;64:669–673.
23. 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.
24. 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.
25. Ingle JN, Ahmann DL, Green SJ, et al. Randomized Clinical Trial of Diethylstilbestrol versus Tamoxifen in Postmenopausal Women with Advanced Breast Cancer. *N Engl J Med* 1981;304:16–21.
26. Roberts R, Croft C, Gold HK, et al. Effect of Propranolol on Myocardial-Infarct Size in a Randomized Blinded Multicenter Trial. *N Engl J Med* 1984;311:218–225.
27. Temple R, Pledger GW. The FDA's Critique of the Anturane Reinfarction Trial. *N Engl J Med* 1980;303:1488–1492.

28. The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the Prevention of Sudden Death after Myocardial Infarction. *N Engl J Med* 1980;302:250–256.
29. The Anturane Reinfarction Trial Research Group. The Anturane Reinfarction Trial: Reevaluation of Outcome. *N Engl J Med* 1982;306:1005–1008.
30. The Canadian Cooperative Study Group. A Randomized Trial of Aspirin and Sulfinpyrazone in Threatened Stroke. *N Engl J Med* 1978;299:53–59.
31. The Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
32. Soran A, Nesbitt L, Mamounas EP, et al. Centralized medical monitoring in phase III clinical trials: the National Surgical Adjuvant Breast and Bowel Project (NSABP) experience. *Clin Trials* 2006;3:478–485.
33. Reboussin D, Espeland MA. The science of web-based clinical trial management. *Clin Trials* 2005;2:1–2.
34. CIBIS-II Investigators and Committees. The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial. *Lancet* 1999;353:9–13.
35. Ambrosius WT, Sink KM, Foy CG, et al. The SPRINT Study Research Group. The design and rationale of a multicenter clinical trial comparing two strategies for control of systolic blood pressure: The Systolic Blood Pressure Intervention Trial (SPRINT). *Clin Trials* 2014;11:532–546.
36. Kjekshus J, Apetrei E, Barrios V, et al. Rosuvastatin in Older Patients with Systolic Heart Failure. *N Engl J Med* 2007;357:2248–2261.
37. 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.
38. 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.
39. The GUSTO Investigators. An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myocardial Infarction. *N Engl J Med* 1993;329:673–682.
40. Detre K, Peduzzi P. The problem of attributing deaths of nonadherers: The VA coronary bypass experience. *Control Clin Trials* 1982;3:355–364.
41. Diggle PJ. Testing for Random Dropouts in Repeated Measurement Data. *Biometrics* 1989;45:1255–1258.
42. Dillman RO, Seagren SL, Propert KJ, et al. A Randomized Trial of Induction Chemotherapy plus High-Dose Radiation versus Radiation Alone in Stage III Non-Small-Cell Lung Cancer. *N Engl J Med* 1990;323:940–945.
43. Dolin R, Reichman RC, Madore HP, et al. A Controlled Trial of Amantadine and Rimantadine in the Prophylaxis of Influenza a Infection. *N Engl J Med* 1982;307:580–584.
44. Heyting A, Tolboom JTBM, Essers JGA. Statistical handling of drop-outs in longitudinal clinical trials. *Statist Med* 1992;11:2043–2061.
45. Hoover DR, Munoz A, Carey V, et al. Using Events from Dropouts in Nonparametric Survival Function Estimation with Application to Incubation of AIDS. *J Am Stat Assoc* 1993;88:37–43.
46. Lagakos SW, Lim LLY, Robins JM. Adjusting for early treatment termination in comparative clinical trials. *Statist Med* 1990;9:1417–1424.
47. Lipid Research Clinics Program. The lipid research clinics coronary primary prevention trial results: I. reduction in incidence of coronary heart disease. *JAMA* 1984;251:351–364.
48. Morgan TM. Analysis of duration of response: A problem of oncology trials. *Control Clin Trials* 1988;9:11–18.
49. Oakes D, Moss AJ, Fleiss JL, et al. Use of Compliance Measures in an Analysis of the Effect of Diltiazem on Mortality and Reinfarction After Myocardial Infarction. *J Am Stat Assoc* 1993;88:44–49.
50. Pizzo PA, Robichaud KJ, Edwards BK, et al. Oral antibiotic prophylaxis in patients with cancer: a double-blind randomized placebo-controlled trial. *J Pediatr* 1983;102:125–133.

51. Pledger GW. Basic statistics: importance of compliance. *Journal of clinical research and pharmacoepidemiology* 1992;6:77–81.
52. Redmond C, Fisher B, Wieand HS. The methodologic dilemma in retrospectively correlating the amount of chemotherapy received in adjuvant therapy protocols with disease-free survival. *Cancer Treat Rep* 1983;67:519–526.
53. Ridout MS, Diggle PJ: Testing for Random Dropouts in Repeated Measurement Data. *Biometrics* 1991;47:1617–1621.
54. Simon R, Makuch RW. A non-parametric graphical representation of the relationship between survival and the occurrence of an event: Application to responder versus non-responder bias. *Statist Med* 1984;3:35–44.
55. Sommer A, Zeger SL. On estimating efficacy from clinical trials. *Statist Med* 1991;10:45–52.
56. The Coronary Drug Project Research Group. The coronary drug project: Initial findings leading to modifications of its research protocol. *JAMA* 1970;214:1303–1313.
57. The Coronary Drug Project Research Group. Influence of Adherence to Treatment and Response of Cholesterol on Mortality in the Coronary Drug Project. *N Engl J Med* 1980;303:1038–1041.
58. Verter J, Friedman L. Adherence measures in the aspirin myocardial infarction study (AMIS). *Control Clin Trials* 1984;5:306.
59. Wilcox RGR, Roland J, Banks D, et al. Randomised trial comparing propranolol with atenolol in immediate treatment of suspected myocardial infarction. *BMJ* 1980;280:885-888.
60. FDA: Guidance for Industry: Non-Inferiority Trials. FDA.
61. Espeland MA, Byington RP, Hire D, et al. Analysis strategies for serial multivariate ultrasonographic data that are incomplete. *Statist Med* 1992;11:1041–1056.
62. The Intermittent Positive Pressure Breathing Trial Group. Intermittent Positive Pressure Breathing Therapy of Chronic Obstructive Pulmonary Disease. A Clinical Trial. *Ann Intern Med* 1983;99:612–620.
63. Conaway MR, Rejeski WJ, Miller ME. Statistical issues in measuring adherence: Methods for incomplete longitudinal data; in Riekert KA, Judith KO, Shumaker SA (eds): The Handbook of Health Behavior Change. New York, Springer Publishing Company, 2008, pp 375–391.
64. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol* 1977;1–38.
65. Efron B. Missing Data, Imputation, and the Bootstrap. *J Am Stat Assoc* 1994;89:463–475.
66. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis, ed 2nd. John Wiley & Sons, 2011.
67. Greenlees JS, Reece WS, Zieschang KD. Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed. *J Am Stat Assoc* 1982;77:251–261.
68. Laird NM. Missing data in longitudinal studies. *Statist Med* 1988;7:305–315.
69. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Wiley, 2002.
70. Little RJA. Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *J Am Stat Assoc* 1995;90:1112–1121.
71. Molenberghs G, Kenward M. Missing Data in Clinical Studies. Wiley, 2007.
72. Rubin DB. Inference and missing data. *Biometrika* 1976;63:581–592.
73. Shao J, Zhong B. Last observation carry-forward and last observation analysis. *Statist Med* 2003;22:2429–2441.
74. O'Kelly M, Ratitch B. Clinical Trials with Missing Data: A Guide for Practitioners. Wiley, 2014.
75. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley, 2009.
76. Steering Committee of the Physicians' Health Study. Final Report on the Aspirin Component of the Ongoing Physicians' Health Study. *N Engl J Med* 1989;321:129–135.
77. 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.

78. Wu MC, Carroll RJ. Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics* 1988;44:175–188.
79. Wu MC, Bailey KR. Estimation and comparison of Changes in the Presence of Informative Right Censoring: Conditional Linear Model. *Biometrics* 1989;45:939–955.
80. Bristow MR, Saxon LA, Boehmer J, et al. Cardiac-Resynchronization Therapy with or without an Implantable Defibrillator in Advanced Chronic Heart Failure. *N Engl J Med* 2004;350:2140–2150.
81. Bresalier RS, Sandler RS, Quan H, et al. Cardiovascular Events Associated with Rofecoxib in a Colorectal Adenoma Chemoprevention Trial. *N Engl J Med* 2005;352:1092–1102.
82. Lagakos SW. Time-to-Event Analyses for Long-Term Treatments—The APPROVe Trial. *N Engl J Med* 2006;355:113–117.
83. Nissen SE. Adverse Cardiovascular Effects of Rofecoxib. *N Engl J Med* 2006;355:203-205.
84. Baron JA, Sandler RS, Bresalier RS, et al. Cardiovascular events associated with rofecoxib: final analysis of the APPROVe trial. *Lancet* 2008;372:1756–1764.
85. Kruskal WH. Some remarks on wild observations. *Technometrics* 1960;2:1–3.
86. Canner PL, Huang YB, Meinert CL. On the detection of outlier clinics in medical and surgical trials: I. Practical considerations. *Control Clin Trials* 1981;2:231–240.
87. Canner PL, Huang YB, Meinert CL. On the detection of outlier clinics in medical and surgical trials: II. Theoretical considerations. *Control Clin Trials* 1981;2:241–252.
88. Dixon WJ. Processing Data for Outliers. *Biometrics* 1953;9:74–89.
89. Grubbs FE. Procedures for detecting outlying observations in samples. *Technometrics* 1969;11:1–21.
90. Anand IS, Carson P, Galle E, et al. Cardiac Resynchronization Therapy Reduces the Risk of Hospitalizations in Patients With Advanced Heart Failure: Results From the Comparison of Medical Therapy, Pacing and Defibrillation in Heart Failure (COMPANION) Trial. *Circulation* 2009;119:969–977.
91. Cannon CP, Braunwald E, McCabe CH, et al. Intensive versus Moderate Lipid Lowering with Statins after Acute Coronary Syndromes. *N Engl J Med* 2004;350:1495–1504.
92. Shepherd J, Cobbe SM, Ford I, et al. Prevention of Coronary Heart Disease with Pravastatin in Men with Hypercholesterolemia. *N Engl J Med* 1995;333:1301–1308.
93. Scandinavian Simvastatin Survival Study Group. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 1994;344:1383–1389.
94. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing, 2013.
95. SAS Institute: SAS/STAT 12.1 User's Guide: Survival Analysis. SAS Institute, 2012.
96. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 2007;334:786.
97. Tomlinson G, Detsky AS. Composite end points in randomized trials: There is no free lunch. *JAMA* 2010;303:267–268.
98. The Look AHEAD Research Group. Look AHEAD (Action for Health in Diabetes): design and methods for a clinical trial of weight loss for the prevention of cardiovascular disease in type 2 diabetes. *Control Clin Trials* 2003;24:610–628.
99. 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.
100. Look AHEAD Research Group. Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes. *N Engl J Med* 2013;369:145–154.
101. Committee of Principal Investigators, World Health Organization. A co-operative trial in the primary prevention of ischaemic heart disease using clofibrate. *Br Heart J* 1978;40:1069–1118.

102. Aspirin Myocardial Infarction Study Research Group. A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 1980;243:661–669.
103. Crager MR. Analysis of Covariance in Parallel-Group Clinical Trials with Pretreatment Baselines. *Biometrics* 1987;43:895–901.
104. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol* 1983;1:710–719.
105. Byar DP. Assessing apparent treatment—covariate interactions in randomized clinical trials. *Statist Med* 1985;4:255–263.
106. Thall PF, Lachin JM. Assessment of stratum-covariate interactions in Cox's proportional hazards regression model. *Statist Med* 1986;5:73–83.
107. Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. *Control Clin Trials* 1989;10:161–175.
108. Weiss GB, Bunce III H, James A. Comparing survival of responders and nonresponders after treatment: A potential source of confusion in interpreting cancer clinical trials. *Control Clin Trials* 1983;4:43–52.
109. Efron B, Feldman D. Compliance as an Explanatory Variable in Clinical Trials. *J Am Stat Assoc* 1991;86:9–17.
110. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Stat Theory Methods* 1991;20:2609–2631.
111. Canner PL. Covariate adjustment of treatment effects in clinical trials. *Control Clin Trials* 1991;12:359–366.
112. Morgan TM, Elashoff RM. Effect of covariate measurement error in randomized clinical trials. *Statist Med* 1987;6:31–41.
113. Canner PL. Further aspects of data analysis. *Control Clin Trials* 1983;4:485-503.
114. Shuster J, van Eys J. Interaction between prognostic factors and treatment. *Control Clin Trials* 1983;4:209–214.
115. Albert JM, DeMets DL. On a model-based approach to estimating efficacy in clinical trials. *Statist Med* 1994;13:2323–2335.
116. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Series B Stat Methodol* 1972;34:187–220.
117. Oye RK, Shapiro MF. Reporting results from chemotherapy trials: Does response make a difference in patient survival? *JAMA* 1984;252:2722–2725.
118. Peto R. Statistical aspects of cancer trials; in Halnan KE (ed): Treatment of Cancer. London, Chapman and Hall, 1982.
119. Gail MH, Simon R. Testing for Qualitative Interactions between Treatment Effects and Patient Subsets. *Biometrics* 1985;41:361–372.
120. Yates F. The Analysis of Multiple Classifications with Unequal Numbers in the Different Classes. *J Am Stat Assoc* 1934;29:51–66.
121. Rosenbaum PR. The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. *J R Stat Soc Ser A* 1984;147:656–666.
122. Egger MJ, Coleman ML, Ward JR, et al. Uses and abuses of analysis of covariance in clinical trials. *Control Clin Trials* 1985;6:12–24.
123. Armitage P, Berry G, Mathews J. Statistical Methods in Medical Research, ed 4th. Malden MA, Blackwell Publishing, 2002.
124. Balke A, Pearl J. Bounds on Treatment Effects From Studies With Imperfect Compliance. *J Am Stat Assoc* 1997;92:1171–1176.
125. Loeys T, Goetghebeur E. A Causal Proportional Hazards Estimator for the Effect of Treatment Actually Received in a Randomized Trial with All-or-Nothing Compliance. *Biometrics* 2003;59:100–105.
126. Sagarin BJ, West SG, Ratnikov A, et al. Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychol Methods* 2014;19:317–333.

127. Ten Have TR, Normand SL, Marcus SM, et al. Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials. *Psychiatr Ann* 2008;38:772.
128. Andersen MP, Frederiksen J, Jnrgensen H, et al. Effect of Alprenolol on Mortality Among Patients With Definite or Suspected Acute Myocardial Infarction. *Lancet* 1979;314:865–868.
129. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064–1069.
130. Baas C, Strackee J, Jones I. Lung Cancer and Month of Birth. *Lancet* 1964;283:47.
131. Bhatt DL, Fox KAA, Hacke W, et al. Clopidogrel and Aspirin versus Aspirin Alone for the Prevention of Atherothrombotic Events. *N Engl J Med* 2006;354:1706–1717.
132. Brookes ST, Whitely E, Egger M, et al. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol* 2004;57:229–236.
133. Collins P, Mosca L, Geiger MJ, et al. Effects of the Selective Estrogen Receptor Modulator Raloxifene on Coronary Outcomes in The Raloxifene Use for the Heart Trial: Results of Subgroup Analyses by Age and Other Factors. *Circulation* 2009;119:922–930.
134. Davies JM. Cancer and Date of Birth. *BMJ* 1963;2:1535.
135. Dijkstra BKS. Origin of carcinoma of the bronchus. *J Natl Cancer Inst* 1963;31:511–519.
136. Furberg CD, Byington RP. What do subgroup analyses reveal about differential response to beta-blocker therapy? The Beta-Blocker Heart Attack Trial experience. *Circulation* 1983;67: I98–101.
137. Furberg CD, Hawkins CM, Lichstein E. Effect of propranolol in postinfarction patients with mechanical or electrical complications. *Circulation* 1984;69:761–765.
138. Goudie RB. The Birthday Fallacy and Statistics of Icelandic Diabetes. *Lancet* 1981;318:1173.
139. Helgason T, Jonasson MR. Evidence For A Food Additive As A Cause Of Ketosis-Prone Diabetes. *Lancet* 1981;318:716–720.
140. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. Biostatistics in clinical medicine. New York, MacMillan, 1983.
141. 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.
142. Kaul S, Diamond GA. Trial and Error: How to Avoid Commonly Encountered Limitations of Published Clinical Trials. *J Am Coll Cardiol* 2-2-2010;55:415–427.
143. Lagakos SW. The Challenge of Subgroup Analyses—Reporting without Distorting. *N Engl J Med* 2006;354:1667–1669.
144. Lee KL, McNeer JF, Starmer CF, et al. Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. *Circulation* 1980;61:508–515.
145. Multicentre International Study. Improvement in prognosis of myocardial infarction by long-term beta-adrenoreceptor blockade using practolol. A multicentre international study. *BMJ* 1975;3:735–740.
146. 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.
147. Packer M, Carson P, Elkayam U, et al. Effect of Amlodipine on the Survival of Patients With Severe Chronic Heart Failure Due to a Nonischemic Cardiomyopathy: Results of the PRAISE-2 Study (Prospective Randomized Amlodipine Survival Evaluation 2). *JACC: Heart Fail* 2013;1:308–314.
148. Pfeffer MA, Jarcho JA. The Charisma of Subgroups and the Subgroups of CHARISMA. *N Engl J Med* 2006;354:1744–1746.
149. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statist Med* 2002;21:2917–2930.

150. Simon R. Patient subsets and variation in therapeutic efficacy. *Br J Clin Pharmacol* 1982;14:473–482.
151. Thackray S, Witte K, Clark AL, Cleland JG. Clinical trials update: OPTIME-CHF, PRAISE-2, ALLHAT. *Euro J Heart Fail* 2000;2:209–212.
152. The ACCORD Study Group. Effects of Intensive Blood-Pressure Control in Type 2 Diabetes Mellitus. *N Engl J Med* 2010;362:1575–1585.
153. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine—Reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007;357:2189–2194.
154. Wedel H, DeMets D, Deedwania P, et al. Challenges of subgroup analyses in multinational clinical trials: experiences from the MERIT-HF trial. *Am Heart J* 2001;142:502–511.
155. Wittes J: On Looking at Subgroups. *Circulation* 2009;119:912–915.
156. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266:93–98.
157. Carson P, Ziesche S, Johnson G, Cohn JN. Racial differences in response to therapy for heart failure: Analysis of the vasodilator-heart failure trials. *J Card Fail* 1999;5:178–187.
158. Cohn JN, Archibald DG, Francis GS, et al. Veterans Administration Cooperative Study on Vasodilator Therapy of Heart Failure: influence of prerandomization variables on the reduction of mortality by treatment with hydralazine and isosorbide dinitrate. *Circulation* 1987;75: IV49-IV54.
159. Franciosa JA, Taylor AL, Cohn JN, et al. African-American Heart Failure Trial (A-HeFT): Rationale, design, and methodology. *J Card Fail* 2002;8:128–135.
160. 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.
161. Hennekens CH, DeMets D. The need for large-scale randomized evidence without undue emphasis on small trials, meta-analyses, or subgroup analyses. *JAMA* 2009;302:2361–2362.
162. Sedgwick P. Randomised controlled trials: subgroup analyses. *BMJ* 2014;349.
163. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer* 1977;35:1.
164. Miller RG. Simultaneous Statistical Inference, ed 2nd. Springer New York, 2011.
165. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988;75:800–802.
166. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 1979;6:65–70.
167. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982;3:345–353.
168. Califf RM. A perspective on the regulation of the evaluation of new antithrombotic drugs. *Am J Cardiol* 1998;82:25P–35P.
169. D'Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statist Med* 2003;22:169–186.
170. Diamond GA, Kaul S. An Orwellian discourse on the meaning and measurement of noninferiority. *Am J Cardiol* 2007;99:284–287.
171. Ellenberg SS, Temple R. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. *Ann Intern Med* 2000;133:464–470.
172. Fleming TR. Current issues in non-inferiority trials. *Statist Med* 2008;27:317-332.
173. Hasselblad V, Kong DF. Statistical methods for comparison to placebo in active-control trials. *Drug Inf J* 2001;35:435–449.
174. James Hung HM, Wang SJ, Tsong Y, et al. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat Med* 2003;22:213–225.
175. Hung HMJ, Wang S, O'Neill R. A Regulatory Perspective on Choice of Margin and Statistical Inference Issue in Non-inferiority Trials. *Biom J* 2005;47:28–36.

176. Kaul S, Diamond GA, Weintraub WS. Trials and tribulations of non-inferiority: the ximelagatran experience. *J Am Coll Cardiol* 2005;46:1986–1995.
177. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of noninferiority trials. *Ann Intern Med* 2006;145:62–69.
178. Kaul S, Diamond GA. Making sense of noninferiority: a clinical and statistical perspective on its application to cardiovascular clinical trials. *Prog Cardiovasc Dis* 2007;49:284–299.
179. Koch A, Röhmel J: The impact of sloppy study conduct on noninferiority studies. *Drug Inf J* 2002;36:3–6.
180. Piaggio G, Elbourne DR, Altman DG, et al., CONSORT Group. Reporting of noninferiority and equivalence randomized trials: An extension of the consort statement. *JAMA* 2006;295:1152–1160.
181. Pocock SJ, Ware JH. Translating statistical findings into plain English. *Lancet* 2009;373:1926–1928.
182. Siegel JP. Equivalence and noninferiority trials. *Am Heart J* 2000;139:S166-S170.
183. SPORTIF Executive Steering Committee for the SPORTIF. Ximelagatran vs warfarin for stroke prevention in patients with nonvalvular atrial fibrillation: A randomized trial. *JAMA* 2005;293:690–698.
184. Cui L, Hung HMJ, Wang SJ. Modification of Sample Size in Group Sequential Clinical Trials. *Biometrics* 1999;55:853–857.
185. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990;335:149–153.
186. DeMets DL. Methods for combining randomized clinical trials: Strengths and limitations. *Statist Med* 1987;6:341–348.
187. Goodman SN. Meta-analysis and evidence. *Control Clin Trials* 1989;10:188–204.
188. Hennekens CH, Buring JE, Hebert PR. Implications of overviews of randomized trials. *Statist Med* 1987;6:397–402.
189. Meinert CL. Meta-analysis: Science or religion? *Control Clin Trials* 1989;10:257–263.
190. Peto R. Why do we need systematic overviews of randomized trials? (Transcript of an oral presentation, modified by the editors). *Statist Med* 1987;6:233–240.
191. Sacks HS, Berrier J, Reitman D, et al. Meta-Analyses of Randomized Controlled Trials. *N Engl J Med* 1987;316:450–455.
192. Simon R. The role of overviews in cancer therapeutics. *Statist Me*d 1987;6:389-393.
193. Yusuf S. Obtaining medically meaningful answers from an overview of randomized clinical trials. *Statist Med* 1987;6:281–286.
194. Cochran WG. Some Methods for Strengthening the Common χ^2 Tests. *Biometrics* 1954;10:417–451.
195. Mantel N, Haenszel W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *J Natl Cancer Inst* 1959;22:719–748.
196. Berlin JA, Laird NM, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. *Statist Med* 1989;8:141–151.
197. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. Wiley, 2011.
198. Brand R, Kragt H. Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statist Med* 1992;11:2077–2082.
199. Carroll RJ, Stefanski LA. Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statist Med* 1994;13:1265–1282.
200. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177-188.
201. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statist Med* 1988;7:889–894.
202. Hedges LV, Olkin I. Statistical Methods for Meta-analysis. Academic Press, 1985.
203. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Statist Med* 1990;9:657–671.
204. Stangl D, Berry DA. Meta-Analysis in Medicine and Health Policy. Taylor & Francis, 2000.

205. Thompson SG. Meta-analysis of clinical trials; in Armitage P, Colton T (eds): Encyclopedia of Biostatistics. New York, Wiley, 1998, pp 2570–2579.
206. Whitehead A. Meta-Analysis of Controlled Clinical Trials. Wiley, 2003.
207. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statist Med* 1991;10:1665–1677.
208. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Wiley, 2008.
209. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med* 2009;151:W–65.
210. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Ann Intern Med* 2009;151:264–269.
211. Baum ML, Anish DS, Chalmers TC, et al. A Survey of Clinical Trials of Antibiotic Prophylaxis in Colon Surgery: Evidence against Further Use of No-Treatment Controls. *N Engl J Med* 1981;305:795–799.
212. Burzotta F, De Vita M, Gu YL, et al. Clinical impact of thrombectomy in acute ST-elevation myocardial infarction: an individual patient-data pooled analysis of 11 trials. *Eur Heart J* 2009;30:2193–2203.
213. Canner PL. Aspirin in coronary heart disease. Comparison of six clinical trials. *Isr J Med Sci*1983;19:413–423.
214. Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence Favoring the Use of Anticoagulants in the Hospital Phase of Acute Myocardial Infarction. *N Engl J Med* 1977;297:1091–1096.
215. Hennekens CH, Buring JE, Sandercock P, et al. Aspirin and other antiplatelet agents in the secondary and primary prevention of cardiovascular disease. *Circulation* 1989;80:749–756.
216. Himel HN, Liberati A, Gelber RD, Chalmers TC. Adjuvant chemotherapy for breast cancer: A pooled estimate based on published randomized control trials. *JAMA* 1986;256:1148–1159.
217. Kotecha D, Holmes J, Krum H, et al. Efficacy of ß blockers in patients with heart failure plus atrial fibrillation: an individual-patient data meta-analysis. *Lancet* 2014;384:2235–2243.
218. May GS, Furberg CD, Eberlein KA, Geraci BJ. Secondary prevention after myocardial infarction: A review of short-term acute phase trials. *Prog Cardiovasc Dis* 1983;25:335–359.
219. Wang PH, Lau J, Chalmers TC. Meta-analysis of effects of intensive blood-glucose control on late complications of type I diabetes. *Lancet* 1993;341:1306–1309.
220. Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Prog Cardiovasc Dis* 1985;27:335–371.
221. Yusuf S, Collins R, Peto R, et al. Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. *Eur Heart J* 1985;6:556–585.
222. Bailey KR: Inter-study differences: How should they influence the interpretation and analysis of results? *Statist Med* 1987;6:351–358.
223. Berlin JA, Begg CB, Louis TA. An Assessment of Publication Bias Using a Sample of Published Clinical Trials. *J Am Stat Assoc* 1989;84:381–392.
224. Chalmers TC, Frank CS, Reitman D. Minimizing the three stages of publication bias. *JAMA* 1990;263:1392–1395.
225. Chalmers TC, Levin H, Sacks HS, et al. Meta-analysis of clinical trials as a scientific discipline. I: Control of bias and comparison with large co-operative trials. *Statist Med* 1987;6:315–325.
226. Collins R, Gray R, Godwin J, Peto R. Avoidance of large biases and large random errors in the assessment of moderate treatment effects: The need for systematic overviews. *Statist Med* 1987;6:245–250.
227. Furberg CD, Morgan TM. Lessons from overviews of cardiovascular trials. *Statist Med* 1987;6:295–303.
228. Furberg CD. Lipid-lowering trials: Results and limitations. *Am Heart J* 1994;128:1304–1308.

229. Goldman L, Feinstein AR. Anticoagulants and Myocardial Infarction. The Problems of Pooling, Drowning, and Floating. *Ann Intern med* 1979;90:92–94.
230. Johnson RT, Dickersin K. Publication bias against negative results from clinical trials: three of the seven deadly sins. *Nat Clin Pract Neurol* 2007;3:590–591.
231. Simes RJ. Confronting publication bias: A cohort design for meta-analysis. *Statist Med* 1987;6:11–29.
232. Thompson SG. Systematic Review: Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;309:1351–1355.
233. Wittes RE. Problems in the medical interpretation of overviews. *Statist Med* 1987;6:269-276.
234. Svilaas T, Vlaar PJ, van der Horst IC, et al. Thrombus Aspiration during Primary Percutaneous Coronary Intervention. *N Engl J Med* 2008;358:557–567.
235. Bavry AA, Kumbhani DJ, Bhatt DL. Role of adjunctive thrombectomy and embolic protection devices in acute myocardial infarction: a comprehensive meta-analysis of randomized trials. *Eur Heart J* 2008;29:2989–3001.
236. Costopoulos C, Gorog DA, Di Mario C, Kukreja N. Use of thrombectomy devices in primary percutaneous coronary intervention: A systematic review and meta-analysis. *Int J Cardiol* 2013;163:229–241.
237. De Luca G, Navarese EP, Suryapranata H. A meta-analytic overview of thrombectomy during primary angioplasty. *Int J Cardiol* 2013;166:606–612.
238. Kumbhani DJ, Bavry AA, Desai MY, et al. Role of Aspiration and Mechanical Thrombectomy in Patients With Acute Myocardial Infarction Undergoing Primary Angioplasty: An Updated Meta-Analysis of Randomized Trials. *J Am Coll Cardiol* 2013;62:1409–1418.
239. Mongeon FP, Bélisle P, Joseph L, et al. Adjunctive Thrombectomy for Acute Myocardial Infarction: A Bayesian Meta-Analysis. *Circ Cardiovasc Interv* 2010;3:6–16.
240. Tamhane U, Chetcuti S, Hameed I, et al. Safety and efficacy of thrombectomy in patients undergoing primary percutaneous coronary intervention for Acute ST elevation MI: A Meta-Analysis of Randomized Controlled Trials. *BMC Cardiovasc Disord* 2010;10:10.
241. Lagerqvist B, Fröbert O, Olivecrona GrK, et al. Outcomes 1 Year after Thrombus Aspiration for Myocardial Infarction. *N Engl J Med* 2014;371:1111–1120.
242. Kumbhani DJ, Bavry AA, Desai MY, et al. Aspiration thrombectomy in patients undergoing primary angioplasty: Totality of data to 2013. *Cathet Cardiovasc Intervent* 2014;84:973–977.
243. Gordon D, Taddei-Peters W, Mascette A, et al. Publication of Trials Funded by the National Heart, Lung, and Blood Institute. *N Engl J Med* 2013;369:1926–1934.
244. Chalmers TC. Randomization of the first patient. *Med Clin North Am* 1975;59:1035–1038.