Chapter 3
What Is the Question?

The planning of a clinical trial depends on the question that the investigator is addressing. The general objective is usually obvious, but the specific question to be answered by the trial is often not stated well. Stating the question clearly and in advance encourages proper design. It also enhances the credibility of the findings. The reliability of clinical trial results derives in part from rigorous prospective definition of the hypothesis. This contrasts with observational studies where the analyses are often exploratory, may be part of an iterative process, and therefore more subject to chance [1]. One would like answers to a number of questions, but the study should be designed with only one major question in mind. This chapter discusses the selection of this primary question and appropriate ways of answering it. In addition, types of secondary and subsidiary questions are reviewed.

The first generation of clinical trials typically compared new interventions to placebo or no treatment on top of best current medical care. They addressed the straight-forward question of whether the new treatment was beneficial, neutral, or harmful compared to placebo or nothing. Since that time, the best medical care has improved dramatically, probably largely due to the contribution of randomized clinical trials (see Chap. 1).

Because of this success in developing beneficial therapies and preventive measures, new design challenges emerged. Prospective trial participants are likely to be on proven therapies. A new intervention is then either added to the existing one or compared against it. If a comparison between active treatments is performed in a clinical practice setting, the studies are often referred to as comparative effectiveness research. (Not all comparative effectiveness research involves clinical trials, but this book will be limited to a discussion of trials.) Due to the lower event rate in patients receiving best known care, whether in add-on trials or comparison trials, the margins for improvement with newer interventions became smaller. This statistical power issue has been addressed in three ways: first, sample sizes have been increased (see Chap. 8); second there has been an increased reliance on composite outcomes; and third, there has been an increased use of surrogate outcomes.

© Springer International Publishing Switzerland 2015

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-3-319-18539-2_3

Another consequence of better treatment was the emergence of trials designed to answer a different type of question. In the past, as noted above, the typical question was: Is the new intervention better, or superior to, no treatment or standard treatment? Now, we frequently ask: Do alternative treatments that may be equal to, or at least no worse than, existing treatments with regard to the primary outcome convey other important advantages in terms of safety, adherence, patient convenience, and/or cost? These trials are often referred to as noninferiority trials and are discussed later in this chapter and in more detail in Chaps. 5, 8, and 18.

Fundamental Point

Each clinical trial must have a primary question. The primary question, as well as any secondary or subsidiary questions, should be carefully selected, clearly defined, and stated in advance.

Selection of the Questions

Primary Question

The primary question should be the one the investigators and sponsors are most interested in answering and that is capable of being adequately answered. It is the question upon which the sample size of the study is based, and which must be emphasized in the reporting of the trial results. The primary question may be framed in the form of testing a hypothesis because most of the time an intervention is postulated to have a particular outcome which, on the average, will be different from (or, in the case of noninferiority trials, not worse than) the outcome in a control group [2]. The outcome may be a clinical event such as improving survival, ameliorating an illness or disease complications, reducing symptoms, or improving quality of life; modifying an intermediate or surrogate characteristic such as blood pressure; or changing a biomarker such as a laboratory value.

Sometimes, trials are designed with more than one primary question. This may be appropriate, depending on the trial design. For example, factorial design trials are specifically conducted to answer more than one question. If done in the context of the usual parallel design trial, statistical adjustments might need to be made to account for the additional question(s) and the sample size made adequate. See Chap. 8 for further discussion of the issue of adjustments in parallel design trials.

Secondary Questions Regarding Benefit

There may also be a variety of subsidiary, or *secondary questions* that are usually related to the primary question. The study may be designed to help address these, or else data collected for the purpose of answering the primary question may also elucidate the secondary questions. They can be of two types. In the first, the response variable is different than that in the primary question. For example, the primary question might ask whether mortality from any cause is altered by the intervention. Secondary questions might relate to incidence of cause-specific death (such as cancer mortality), incidence of non-fatal renal failure, or incidence of stroke. Many investigators also assess patient-reported outcomes such as health-related quality of life (see Chap. 13).

The second type of secondary question relates to *subgroup hypotheses*. For example, in a study of cancer therapy, the investigator may want to look specifically at people by gender, age, stage of disease at entry into the trial or by presence or absence of a particular biomarker or genetic marker. Such subsets of people in the intervention group can be compared with similar people in the control group. Subgroup hypotheses should be 1) specified before data collection begins, 2) based on reasonable expectations, and 3) limited in number. In any event, the number of participants in most subgroups is usually too small to prove or disprove a subgroup hypothesis. One should not expect significant differences in subgroup unless the trial was specifically designed to detect them. Failure to find significant differences should not be interpreted to mean that they do not exist. Investigators should exercise caution in accepting subgroup results, especially when the overall trial results are not significant. A survey of clinical trialists indicated that inappropriate subgroup analyses were considered one of the two major sources of distortion of trial findings [3]. Generally, the most useful reasons for considering subgroups are to examine consistency of results across pre-defined subgroups and to create hypotheses that can be tested in future trials and meta-analyses.

There has been recognition that certain subgroups of people have not been adequately represented in clinical research, including clinical trials [4]. In the United States, this has led to requirements that women and minority populations be included in appropriate numbers in trials supported by federal government agencies [5]. The debate is whether the numbers of participants of each sex and racial/ethnic group must be adequate to answer the key questions that the trial addresses, or whether there must merely be adequate diversity of people. Many trials are international in scope. Whether one should examine outcome data by country or region has been debated [6]. Are observed differences in intervention effect by geographic region true or due to the play of chance [7, 8]? One might expect that culture, medical care system, genetic makeup, and other factors could affect the magnitude, or even existence of benefit from a new intervention. But, as has been noted [9, 10], the design and size of the trial should be driven by reasonable expectations that the intervention will or will not operate materially differently among the various subsets of participants. If such variability is expected,

it is appropriate to design the trial to detect those differences. If not, adequate diversity with the opportunity to examine subgroup responses at the end of the trial (and conduct additional research if necessary) is more appropriate.

Secondary questions raise several trial methodological issues; for example, if enough statistical tests are done, a few will be significant by chance alone when there is no true intervention effect. An example was provided by the Second International Study of Infarct Survival (ISIS-2), a factorial design trial of aspirin and streptokinase in patients with acute myocardial infarction [11]. To illustrate the hazards of subgroup analyses, the investigators showed that participants born under the Gemini or Libra astrological birth signs had a somewhat greater incidence of vascular and total mortality on aspirin than on no aspirin, whereas for all other signs, and overall, there was an impressive and highly significant benefit from aspirin. Therefore, when a number of tests are carried out, results should be interpreted cautiously as they may well be due to chance. Shedding light or raising new hypotheses, and perhaps conducting meta-analyses, are more proper outcomes of these analyses than are conclusive answers. See Chap. 18 for further discussion of subgroup and meta-analyses.

Both primary and secondary questions should be important and relevant scientifically, medically, or for public health purposes. Participant safety and well-being must always be considered in evaluating importance. Potential benefit and risk of harm should be looked at by the investigator, as well as by local ethical review committees, and often, independent data monitoring committees.

Questions Regarding Harm

Important questions that can be answered by clinical trials concern adverse effects of or reactions to therapy (Chap. 12). Here, unlike the primary or secondary questions, it is not always possible to specify in advance the questions to be answered. What adverse effects might occur, and their severity, may be unpredictable. Furthermore, rigorous, convincing demonstration of serious toxicity is usually not achieved, because it is generally thought unethical to continue a study to the point at which a drug has been conclusively shown to be more harmful than beneficial [12–14]. Investigators traditionally monitor a variety of laboratory and clinical measurements, look for possible adverse events, and compare these in the intervention and control groups. Some of the most serious adverse effects, however, are rare and do not occur commonly enough to be detected reliably in clinical trials. Statistical significance and the previously mentioned problem of multiple response variables become secondary to clinical judgment and participant safety. While this will lead to the conclusion that some purely chance findings are labeled as adverse effects, responsibility to the participants requires a conservative attitude toward safety monitoring, particularly if an alternative therapy is available. Trials have been stopped early for less than statistically convincing evidence of adverse effects [15–17]. In such cases, only other trials of the identical or related

interventions noting the same adverse effect (as were the situations for these examples of antiarrhythmic therapy in people with heart disease, beta carotene in people at high risk of lung cancer, and an angiotensin-converting enzyme inhibitor in acute myocardial infarction) or convincing nonclinical studies will provide irrefutable evidence that the adverse finding is true. In the last case cited, other studies contradicted the finding.

Ancillary Questions

Often a clinical trial can be used to answer questions which do not bear directly on the intervention being tested, but which are nevertheless of interest. The structure of the trial and the ready access to participants may make it the ideal vehicle for such investigations. Large trials, in particular, create databases that offer opportunities to better understand the disease or condition, treatment, predictors of outcomes, and new hypotheses that can be tested. The Group Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO-1) trial [18] provides an example of use of a dataset that yielded over 100 subsequent publications, including one identifying predictors of mortality [19]. The Assessment of Pexelizumab in Acute Myocardial Infarction (APEX AMI) trial [20] found no benefit from the complement inhibitor, pexelizumab, but so far, over 50 manuscripts regarding primary angioplasty in acute ST-elevation myocardial infarction have been published.

Clinical trials can also be used to examine issues such as how the intervention works. A small group of participants might undergo mechanistic studies (as long as they are not unduly burdensome or invasive). In the Studies of Left Ventricular Dysfunction (SOLVD) [21], the investigators evaluated whether an angiotensin converting enzyme inhibitor would reduce mortality in symptomatic and asymptomatic subjects with impaired cardiac function. In selected participants, special studies were done with the objective of getting a better understanding of the disease process and of the mechanisms of action of the intervention. These substudies did not require the large sample size of the main studies (over 6,000 participants). Thus, most participants in the main trials had a relatively simple and short evaluation and did not undergo the expensive and time-consuming procedures or interviews demanded by the substudies. This combination of a rather limited assessment in many participants, designed to address an easily monitored response variable, and detailed measurements in subsets, can be extremely effective. An angiographic substudy in the GUSTO trial helped explain how accelerated alteplase treatment resulted in more effective coronary perfusion [22]. The improved survival appeared to be fully explained by this impact on reperfusion [23]. In the Harmonizing Outcomes with Revascularization and Stents in Acute Myocardial Infarction (HORIZONS-AMI) trial [24], lower rates of bleeding with bivalirudin compared with unfractionated heparin plus a glycoprotein IIb/IIIa inhibitor appeared to explain only part of the lower subsequent mortality in the bivalirudin group [25].

Exploratory genetic studies are commonly conducted to examine possible mechanisms of action of the intervention. Genetic variants of the cytochrome P450 CYP2C19 metabolic pathway of clopidogrel were related to the level of the active metabolite and reduction in platelet aggregation for participants treated with clopidogrel in the database from the Trial to Assess Improvement in Therapeutic Outcomes by Optimizing Platelet Inhibition with Prasugrel-Thrombolysis in Myocardial Infarction (TRITON-TIMI) [26].

Kinds of Trials

Trials with Extensive Data Collection vs. Large, Simple

Traditionally, most trials of new interventions have collected extensive information about participants, have detailed inclusion and exclusion criteria, involve considerable quality assurance measures, and assess many, carefully measured outcomes. These sorts of trials, although they address major questions and are well-conducted, are quite expensive and often very time-consuming. Therefore, given the needed resources, trial sponsors can afford to address only some of the many important questions can be answered, often in limited kinds of participants and clinical settings.

As discussed by Tricoci et al. [27] with respect to clinical practice guidelines in cardiology, but undoubtedly similar in other medical fields, many of these guidelines are based on inadequate data. One of the rationales for large, simple clinical trials is that they can provide data relevant to clinical practice, since they are typically conducted in practice settings [28]. The general idea is that for common conditions, and important outcomes, such as total mortality, even modest benefits of intervention, particularly interventions that are easily implemented in a large population, are important. Because an intervention is likely to have similar effects (or at least effects that trend in the same direction) in most participants, extensive characterization of people at entry may be unnecessary. The study must have unbiased allocation of participants to intervention or control and unbiased and reasonably complete ascertainment of outcomes. Sufficiently large numbers of participants are more important in providing the statistical power necessary to answer the question(s) than careful attention to quality and completeness of data. This model depends upon a relatively easily administered intervention, brief forms, and an easily ascertained outcome, such as a fatal or unambiguous nonfatal event. Neither the trials that collect extensive information nor the simple ones are better. Rather, both types are essential. The proper design depends on the condition being studied, the nature of the question, and the kind of intervention.

Superiority vs. Noninferiority Trials

As mentioned in the introduction to this chapter, traditionally, most trials were designed to establish whether a new intervention on top of usual or standard care was superior to that care alone (or that care plus placebo). If there were no effective treatments, the new intervention was compared to just placebo. As discussed in Chap. 8, these trials were generally two-sided. That is, the trial was designed to see whether the new intervention was better or worse than the control.

With the development of effective therapies, many trials have been designed to demonstrate that a new intervention is not worse than the intervention previously shown to be beneficial, i.e., an active control, by some prespecified amount. As noted earlier, the motivation for such a question is that the new intervention might not be better than standard treatment on the primary or important secondary outcomes, but may be less toxic, more convenient, less invasive and/or have some other attractive feature, including lower cost. The challenge is to define what is meant by "not worse than." This has been referred to as the "margin of indifference," or δ, meaning that if the new intervention is not less effective than this margin, its use might be of value given the other features. In the analysis of this design, the 95% upper confidence limit would need to be less than this margin in order to claim noninferiority. Defining δ is challenging and will be discussed in Chap. 5.

The question in a noninferiority trial is different than in a superiority trial and affects both the design and conduct of the trial. For example, in the superiority trial, poor adherence will lead to a decreased ability, or power, to detect a meaningful difference. For a noninferiority trial, poor adherence will diminish real and important differences and bias the results towards a noninferiority claim. Thus, great care must be taken in defining the question, the sensitivity of the outcome measures to the intervention being evaluated, and the adherence to the intervention during the conduct of the trial.

Comparative Effectiveness Trials

As mentioned, major efforts are being devoted to conducting comparative effectiveness research. Although comparative effectiveness studies can be of various sorts, encompassing several kinds of clinical research, we will limit our discussion to clinical trials. Much medical care has not been rigorously evaluated, meaning that trials comparing ongoing preventive and treatment approaches are needed. And of course, when new interventions are developed, they must be compared against existing therapy. Additionally, the increasing cost burden of medical care means that even if several treatments are equally effective, we need to consider factors such as cost, tolerability, and ease of administration. Therefore, comparative effectiveness trials are commonly of the noninferiority sort.

Much of the literature on comparative effectiveness research advocates conducting the studies in usual practice settings (often called pragmatic trials) [29, 30] (see Chap. 4). Because these trials are conducted in clinical practice settings, they must be relatively simple, demanding little in the way of effort to screen and assess outcomes. The goal is to compare two interventions, both of which are considered standard care.

Intervention

When the question is conceived, investigators, at the very least have in mind a class or type of intervention. More commonly, they know the precise drug, procedure, or lifestyle modification they wish to study. In reaching such a decision, they need to consider several aspects.

First, the potential benefit of the intervention must be maximized, while possible harm is kept to a minimum. Thus, dose of drug or intensity of rehabilitation and frequency and route of administration are key factors that need to be determined. Can the intervention or intervention approach be standardized, and remain reasonably stable over the duration of the trial? Investigators must also decide whether to use a single drug, biologic, or device, fixed or adjustable doses of drugs, sequential drugs, or drug or device combinations. Devices in particular undergo frequent modifications and updates. Investigators need to be satisfied that any new version that appears during the course of the trial functions sufficiently similarly in important ways to the older versions so that combining data from the versions would be appropriate. Of course, an investigator can use only the version available at the onset of the trial (if it is still obtainable), but the trial will then be criticized for employing the outdated version. For example, coronary stents have evolved and the newer ones have lower risk of stent thrombosis [31]. This development may have altered their relative effectiveness vs. bypass surgery, therefore trials that continued to use the older versions of the stents have little credibility.

Sometimes, it is not only the active intervention, but other factors that apply. In gene transfer studies, the nature of the vector, as well as the actual gene, may materially affect the outcome, particularly when it comes to adverse effects. If the intervention is a procedure, other considerations must be considered. Surgical and other procedures or techniques are frequently modified and some practitioners are more skilled than others. Investigators need to think about learning curves, and at what point someone has sufficient skill to perform the intervention.

Not only the nature of the intervention, but what constitutes the control group regimen must also be considered for ethical reasons, as discussed in Chap. 2, and study design reasons, as discussed in Chap. 5.

Second, the availability of the drug or device for testing needs to be determined. If it is not yet licensed, special approval from the regulatory agency and cooperation or support by the manufacturer are required (see Chap. 22).

Third, investigators must take into account design aspects, such as time of initiation and duration of the intervention, need for special tests or laboratory facilities, and the logistics of blinding in the case of drug studies. Certain kinds of interventions, such as surgical procedures, device implantation, vaccines, and gene transfer may have long-term or even life-long effects. Therefore, investigators might need to incorporate plans for long-term assessment. There had been reports that drug-eluting stents, used in percutaneous coronary intervention, perhaps had a greater likelihood of restenosis than bare-metal stents [32, 33]. Follow-up studies seemed to assuage these concerns [34]. Nevertheless, investigators must consider incorporating plans for long-term assessment. Problems with metal-on-metal hip replacements were only uncovered years after many had been implanted [35, 36]. The rubbing of the metal ball against the metal cup causes metal particles to wear away, possibly leading to both local and systemic adverse effects.

Response Variables

Kinds of Response Variables

Response variables are outcomes measured during the course of the trial, and they define and answer the questions. A response variable may be total mortality, death from a specific cause, incidence of a disease, a complication or specific adverse effect, symptomatic relief, quality of life, a clinical finding, a laboratory measurement, or the cost and ease of administering the intervention. If the primary question concerns total mortality, the occurrence of deaths in the trial clearly answers the question. If the primary question involves severity of arthritis, on the other hand, extent of mobility or a measure of freedom from pain may be reasonably good indicators. In other circumstances, a specific response variable may only partially reflect the overall question. As seen from the above examples, the response variable may show a change from one discrete state (living) to another (dead), from one discrete state to any of several other states (changing from one stage of disease to another) or from one level of a continuous variable to another. If the question can be appropriately defined using a continuous variable, the required sample size may be reduced (Chap. 8). However, the investigator needs to be careful that this variable and any observed differences are clinically meaningful and relevant and that the use of a continuous variable is not simply a device to reduce sample size.

In general, a single response variable should be identified to answer the primary question. If more than one are used, the probability of getting a nominally significant result by chance alone is increased (Chap. 18). In addition, if the different response variables give inconsistent results, interpretation becomes difficult. The investigator would then need to consider which outcome is most important, and explain why the others gave conflicting results. Unless she has made the

determination of relative importance prior to data collection, her explanations are likely to be unconvincing.

Although the practice is not advocated, there may be circumstances when more than one "primary" response variable needs to be looked at. This may be the case when an investigator truly cannot decide which of several response variables relates most closely to the primary question. Ideally, the trial would be postponed until this decision can be made. However, overriding concerns, such as increasing use of the intervention in general medical practice, may compel her to conduct the study earlier. In these circumstances, rather than arbitrarily selecting one response variable which may, in retrospect, turn out to be suboptimal or even inappropriate, investigators prefer to list several "primary" outcomes. An old example is the Urokinase Pulmonary Embolism Trial [37], where lung scan, arteriogram and hemodynamic measures were given as the "primary" response variables in assessing the effectiveness of the agents urokinase and streptokinase. Chapter 8 discusses the calculation of sample size when a study with several primary response variables is designed.

Commonly, investigators prepare an extensive list of secondary outcomes, allowing them to claim that they "prespecified" these outcomes when one or more turn out to reach nominally significant differences. Although prespecification provides some protection against accusations that the findings were data-derived, a long list does not protect against likely play of chance. Far better is a short list of outcomes that are truly thought to be potentially affected by the intervention. *Combining events* to make up a response variable might be useful if any one event occurs too infrequently for the investigator reasonably to expect a significant difference without using a large number of participants. In answering a question where the response variable involves a combination of events, only *one event per participant* should be counted. That is, the analysis is by participant, not by event.

One kind of combination response variable involves two kinds of events. This has been termed a *composite outcome*. It must be emphasized, however, that the composite outcome should be capable of meaningful interpretation such as where all components are related through a common underlying condition or respond to the same presumed mechanism of action of the agent. In a study of heart disease, combined events might be death from coronary heart disease plus nonfatal myocardial infarction. This is clinically meaningful since death from coronary heart disease and nonfatal myocardial infarction might together represent a measure of serious coronary heart disease. Unfortunately, as identified in a survey of 40 trials using composite outcomes by Cordoba et al. [38], there was considerable lack of clarity as to how components were combined and results reported. Difficulties in interpretation can arise if the results of each of the components in such a response variable are inconsistent [39]. In the Physicians' Health Study report of aspirin to prevent cardiovascular disease, there was no difference in mortality, a large reduction in myocardial infarction, and an increase in stroke, primarily hemorrhagic [40]. In this case, cardiovascular mortality was the primary response variable, rather than a combination. If it had been a combination, the interpretation of the results would have been even more difficult than it was [41]. Even more troublesome is

the situation where one of the components in the combination response variable is far less serious than the others. For example, if occurrence of angina pectoris or a revascularization procedure is added, as is commonly done, interpretation can be problematic. Not only are these less serious than cardiovascular death or myocardial infarction, they often occur more frequently. Thus, if overall differences between groups are seen, are these results driven primarily by the less serious components? What if the results for the more serious components (e.g., death) trend in the opposite directions? This is not just theoretical. For example, the largest difference between intervention and control in the Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) trial was seen in the least serious of the four components; the one that occurred most often in the control group [42]. A survey of published trials in cardiovascular disease that used composite response variables showed that half had major differences in both importance and effect sizes of the individual components [43]. Those components considered to be most important had, on average, smaller benefits than the more minor ones. See Chap. 18 for a discussion of analytic and interpretation issues if the components of the composite outcome go in different directions or have other considerable differences in the effect size.

When this kind of combination response variable is used, the rules for interpreting the results and for possibly making regulatory claims about individual components should be established in advance. A survey of the cardiovascular literature found that the use of composite outcomes (often with three or four components) is common, and the components vary in importance [44]. One possible approach is to require that the most serious individual components show the same trend as the overall result. Some have suggested giving each component weights, depending on the seriousness [45, 46]. However, this may lead to trial results framed as unfamiliar scores that are difficult to interpret by clinicians. Although it has sample size implications, it is probably preferable to include in the combined primary response variable only those components that are truly serious, and to assess the other components as secondary outcomes. If an important part of a composite outcome goes in the wrong direction, as occurred with death in the Sodium-Hydrogen Exchange Inhibition to Prevent Coronary Events in Acute Cardiac Conditions (EXPEDITION) trial [47], even benefit in the composite outcome (death or myocardial infarction), is insufficient to conclude that the intervention (in this case, sodium-hydrogen exchange inhibition by means of cariporide during coronary artery bypass graft surgery) should be used. Adding to the concern was an adverse trend for cerebrovascular events.

Another kind of combination response variable involves multiple events of the same sort. Rather than simply asking whether an event has occurred, the investigator can look at the frequency with which it occurs. This may be a more meaningful way of looking at the question than seeking a yes-no outcome. For example, frequency of recurrent transient ischemic attacks or epileptic seizures within a specific follow-up period might comprise the primary response variable of interest. Simply adding up the number of recurrent episodes and dividing by the number of participants in each group in order to arrive at an average would be improper.

Multiple events in an individual are not independent, and averaging gives undue weight to those with more than one episode. One approach is to compare the number of participants with none, one, two, or more episodes; that is, the distribution of the number of episodes, by individual.

Sometimes, study participants enter a trial with a condition that is exhibited frequently. For example, they may have had several episodes of transient atrial fibrillation in the previous weeks or may drink alcohol to excess several days a month. Trial eligibility criteria may even require a minimum number of such episodes. A trial of a new treatment for alcohol abuse may require participants to have at least six alcoholic drinks a day for at least 7 days over the previous month. The investigator needs to decide what constitutes a beneficial response. Is it complete cessation of drinking? Reducing the number of drinks to some fixed level (e.g., no more than two on any given day)? Reducing alcohol intake by some percent, and if so, what percent? Does this fixed level or percent differ depending on the intake at the start of the trial? Decisions must be made based on knowledge of the disease or condition, the kind of intervention and the expectations of how the intervention will work. The clinical importance of improvement versus complete "cure" must also be considered.

Specifying the Question

Regardless of whether an investigator is measuring a primary or secondary response variable, certain rules apply. First, she should define and record the questions in advance, being as specific as possible. She should not simply ask, "Is *A* better than *B*?" Rather, she should ask, "In population *W* is drug *A* at daily dose *X* more efficacious in improving *Z* by *Q* amount over a period of time *T* than drug *B* at daily dose *Y*?" Implicit here is the magnitude of the difference that the investigator is interested in detecting. Stating the questions and response variables in advance is essential for planning of study design and calculation of sample size. As shown in Chap. 8, sample size calculation requires specification of the response variables as well as estimates of the intervention effect. In addition, the investigator is forced to consider what she means by a successful intervention. For example, does the intervention need to reduce mortality by 10 or 25% before a recommendation for its general use is made? Since such recommendations also depend on the frequency and severity of adverse effects, a successful result cannot be completely defined beforehand. However, if a 10% reduction in mortality is clinically important, that should be stated, since it has sample size implications. Specifying response variables and anticipated benefit in advance also eliminates the possibility of the legitimate criticism that can be made if the investigator looked at the data until she found a statistically significant result and then decided that *that* response variable was what she really had in mind all the time

Second, the primary response variable must be capable of being assessed in all participants. Selecting one response variable to answer the primary question in

some participants, and another response variable to answer the same primary question in other participants is not a legitimate practice. It implies that each response variable answers the question of interest with the same precision and accuracy; i.e., that each measures exactly the same thing. Such agreement is unlikely. Similarly, response variables should be measured in the same way for all participants. Measuring a given variable by different instruments or techniques implies that the instruments or techniques yield precisely the same information. This rarely, if ever, occurs. If response variables can be measured only one way in some participants and another way in other participants, two separate studies are actually being performed, each of which is likely to be too small.

Third, unless there is a combination primary response variable in which the participant remains at risk of having additional events, participation generally ends when the primary response variable occurs. "Generally" is used here because, unless death is the primary response variable, the investigator may well be interested in certain events, including adverse events, subsequent to the occurrence of the primary response variable. These events will not change the analysis of the primary response variable but may affect the interpretation of results. For example, deaths taking place after a nonfatal primary response variable has already occurred, but before the official end of the trial as a whole, may be of interest. On the other hand, if a secondary response variable occurs, the participant should remain in the study (unless, of course, it is a fatal secondary response variable). He must continue to be followed because he is still at risk of developing the primary response variable. A study of heart disease may have, as its primary question, death from coronary heart disease and, as a secondary question, incidence of nonfatal myocardial infarction. If a participant suffers a nonfatal myocardial infarction, this counts toward the secondary response variable. However, he ought to remain in the study for analytic purposes and be at risk of developing the primary response variable and of having other adverse events. This is true whether or not he is continued on the intervention regimen. If he does not remain in the study for purposes of analysis of the primary response variable, bias may result. (See Chap. 18 for further discussion of participant withdrawal.)

Fourth, response variables should be capable of unbiased assessment. Truly double-blind studies have a distinct advantage over other studies in this regard. If a trial is not double-blind (Chap. 7), then, whenever possible, response variable assessment should be done by people who are not involved in participant follow-up and who are blinded to the identity of the study group of each participant. Independent reviewers are often helpful. Of course, the use of blinded or independent reviewers does not entirely solve the problem of bias. Unblinded investigators sometimes fill out forms and the participants may be influenced by the investigators. This may be the case during a treadmill exercise performance test, where the impact of the person administering the test on the results may be considerable. Some studies arrange to have the intervention administered by one investigator and response variables evaluated by another. Unless the participant is blinded to his group assignment (or otherwise unable to communicate), this procedure is also vulnerable to bias. One solution to this dilemma is to use only "hard," or objective,

response variables (which are unambiguous and not open to interpretation, such as total mortality or some imaging or laboratory measures read by someone blinded to the intervention assignment). This assumes complete and honest ascertainment of outcome. Double-blind studies have the advantage of allowing the use of softer response variables, since the risk of assessment bias is minimized.

Fifth, it is important to have response variables that can be ascertained as completely as possible. A hazard of long-term studies is that participants may fail to return for follow-up appointments. If the response variable is one that depends on an interview or an examination, and participants fail to return for follow-up appointments information will be lost. Not only will it be lost, but it may be differentially lost in the intervention and control groups. Death or hospitalizations are useful response variables because the investigator can usually ascertain vital status or occurrence of a hospital admission, even if the participant is no longer active in a study. However, only in a minority of clinical trials are they appropriate.

Sometimes, participants withdraw their consent to be in the trial after the trial has begun. In such cases, the investigator should ascertain whether the participant is simply refusing to return for follow-up visits but is willing to have his data used, including data that might be obtained from public records; is willing to have only data collected up to the time of withdrawal used in analyses; or is asking that all of his data be deleted from the study records.

All clinical trials are compromises between the ideal and the practical. This is true in the selection of primary response variables. The most objective or those most easily measured may occur too infrequently, may fail to define adequately the primary question, or may be too costly. To select a response variable which can be reasonably and reliably assessed and yet which can provide an answer to the primary question requires judgment. If such a response variable cannot be found, the wisdom of conducting the trial should be re-evaluated.

Biomarkers and Surrogate Response Variables

A common criticism of clinical trials is that they are expensive and of long duration. This is particularly true for trials which use the occurrence of clinical events as the primary response variable. It has been suggested that response variables which are continuous in nature might substitute for the binary clinical outcomes. Thus, instead of monitoring cardiovascular mortality or myocardial infarction an investigator could examine progress of atherosclerosis by means of angiography or ultrasound imaging, or change in cardiac arrhythmia by means of ambulatory electrocardiograms or programmed electrical stimulation. In the cancer field, change in tumor size might replace mortality. In AIDS trials, change in CD-4 lymphocyte level has been used as a response to treatment instead of incidence of AIDS in HIV positive patients or mortality. Improved bone mineral density has been used as a surrogate for reduction in fractures.

A rationale for use of these "surrogate response variables" is that since the variables are continuous, the sample size can be smaller and the study less expensive than otherwise. Also, changes in the variables are likely to occur before the clinical event, shortening the time required for the trial. Wittes et al. [48] discuss examples of savings in sample size by the use of surrogate response variables.

It has been argued that in the case of truly life-threatening diseases (e.g., AIDS in its early days, certain cancers, serious heart failure), clinical trials should not be necessary to license a drug or other intervention. Given the severity of the condition, lesser standards of proof should be required. If clinical trials are done, surrogate response variables ought to be acceptable, as speed in determining possible benefit is crucial. Potential errors in declaring an intervention useful may therefore not be as important as early discovery of a truly effective treatment.

Even in such instances, however, one should not uncritically use surrogate endpoints [49, 50]. It was known for years that the presence of ventricular arrhythmias correlated with increased likelihood of sudden death and total mortality in people with heart disease [51], as it was presumably one mechanism for the increased mortality. Therefore, it was common practice to administer antiarrhythmic drugs with the aim of reducing the incidence of sudden cardiac death [52, 53]. The Cardiac Arrhythmia Suppression Trial demonstrated, however, that three drugs that effectively treated ventricular arrhythmias were not only ineffective in reducing sudden cardiac death, but actually caused increased mortality [54, 15].

A second example concerns the use of inotropic agents in people with heart failure. These drugs had been shown to improve exercise tolerance and other symptomatic manifestations of heart failure [55]. It was expected that mortality would also be reduced. Unfortunately, clinical trials subsequently showed that mortality was increased [56, 57].

Another example from the cardiovascular field is the Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events (ILLUMINATE). In this trial, the combination of torcetrapib and atorvastatin was compared with atorvastatin alone in people with cardiovascular disease or diabetes. Despite the expected impressive and highly statistically significant increase in HDL-cholesterol and decrease in LDL-cholesterol in the combination group, there was an increase in all-cause mortality and major cardiovascular events [58]. Thus, even though it is well-known that lowering LDL-cholesterol (and possibly increasing HDL-cholesterol) can lead to a reduction in coronary heart disease events, some interventions might have unforeseen adverse consequences. Recent studies looking at the raising of HDL-cholesterol have also been disappointing, despite the theoretical grounds to expect benefit [59]. The Atherothrombosis Intervention in Metabolic Syndrome with Low HDL/High Triglycerides and Impact on Global Health Outcomes (AIM-HIGH) trial [60] and the Second Hearth Protection Study (HPS-2 THRIVE) [61] did not reduce cardiovascular outcomes in the context of lowering LDL-cholesterol.

It was noted that the level of CD-4 lymphocytes in the blood is associated with severity of AIDS. Therefore, despite some concerns [62] a number of clinical trials used change in CD-4 lymphocyte concentration as an indicator of disease status.

If the level rose, the drug was considered to be beneficial. Lin et al., however, argued that CD-4 lymphocyte count accounts for only part of the relationship between treatment with zidovudine and outcome [63]. Choi et al. came to similar conclusions [64]. In a trial comparing zidovudine with zalcitabine, zalcitabine was found to lead to a slower decline in CD-4 lymphocytes than did zidovudine, but had no effect on the death rate from AIDS [65]. Also troubling were the results of a large trial which, although showing an early rise in CD-4 lymphocytes, did not demonstrate any long-term benefit from zidovudine [66]. Whether zidovudine or another treatment was, or was not, truly beneficial is not the issue here. The main point is that the effect of a drug on a surrogate endpoint (CD-4 lymphocytes) is not always a good indicator of clinical outcome. This is summarized by Fleming, who noted that the CD-4 lymphocyte count showed positive results in seven out of eight trials where clinical outcomes were also positive. However, the CD-4 count was also positive in six out of eight trials in which the clinical outcomes were negative [50].

Similar seemingly contradictory results have been seen with cancer clinical trials. In trials of 5-fluorouracil plus leucovorin compared with 5-fluorouracil alone, the combination led to significantly better tumor response, but no difference in survival [67]. Fleming cites other cancer examples as well [50]. Sodium fluoride, because of its stimulation of bone formation, was widely used in the treatment of osteoporosis. Despite this, it was found in a trial in women with postmenopausal osteoporosis to increase bone fragility [68].

These examples do not mean that surrogate response variables should never be used in clinical trials. Nevertheless, they do point out that they should only be used after considering the advantages and disadvantages, recognizing that erroneous conclusions about interventions might occasionally be reached.

Prentice has summarized two key criteria that must be met if a surrogate response variable is to be useful [69]. First, the surrogate must correlate with the true clinical outcome, which most proposed surrogates would likely do. Second, for a surrogate to be valid, it must capture the full effect of the intervention. For example, a drug might lower blood pressure or serum LDL-cholesterol, but as in the ILLUMINATE trial example, have some other deleterious effect that would negate any benefit or even prove harmful.

Another factor is whether the surrogate variable can be assessed accurately and reliably. Is there so much measurement error that, in fact, the sample size requirement increases or the results are questioned? Additionally, will the evaluation be so unacceptable to the participant that the study will become infeasible? If it requires invasive techniques, participants may refuse to join the trial, or worse, discontinue participation before the end. Measurement can require expensive equipment and highly trained staff, which may, in the end, make the trial more costly than if clinical events are monitored. The small sample size of surrogate response variable trials may mean that important data on safety are not obtained [70]. Finally, will the conclusions of the trial be accepted by the scientific and medical communities? If there is insufficient acceptance that the surrogate variable reflects clinical outcome, in spite of the investigator's conviction, there is little point in using such variables.

Many drugs have been approved by regulatory agencies on the basis of surrogate response variables, including those that reduce blood pressure and blood sugar. In the latter case, though, the Food and Drug Administration now requires new diabetes drugs to show that cardiovascular events are not increased [71]. We think that, except in rare instances, whenever interventions are approved by regulatory bodies on the basis of surrogate response variables, further clinical studies with clinical outcomes should be conducted afterward. As discussed by Avorn [72], however, this has not always been the case. He cites examples not only of major adverse effects uncovered after drugs were approved on the basis of surrogate outcomes, but lack of proven clinical benefit. In all decisions regarding approval, the issues of biologic plausibility, risk, benefits, and history of success must be considered.

When are surrogate response variables useful? The situation of extremely serious conditions has been mentioned. Particularly, when serious conditions are also rare, it may be difficult or even impossible to obtain enough participants to use a clinical outcome. We may be forced to rely on surrogate outcomes. Other than those situations, surrogate response variables are useful in early phase development studies, as an aid in deciding on proper dosage and whether the anticipated biologic effects are being achieved. They can help in deciding whether, and how best, to conduct the late phase trials which almost always should employ clinical response variables.

Changing the Question

Occasionally, investigators want to change the primary response variable partway through a trial. Reasons for this might be several, but usually it is because achieving adequate power for the original primary response variable is no longer considered feasible. The event rate might be less than expected, and even extension of the trial might not be sufficient by itself or might be too expensive. The Look AHEAD (Action for Health in Diabetes) trial was designed to see if weight loss in obese or overweight people with type 2 diabetes would result in a reduction in cardiovascular disease. The investigators were confronted with a much lower than expected rate of the primary outcome during the course of the trial, and after 2 years, the data monitoring board recommended expanding the primary outcome. It was changed from a composite of cardiovascular death, myocardial infarction, and stroke to one including hospitalization for angina. In addition, the duration of follow-up was lengthened [73]. As discussed in Chap. 10, recruitment of participants might be too slow to reach the necessary sample size. The Prevention of Events with Angiotensin Converting Enzyme Inhibition (PEACE) trial was seeking 14,100 patients with coronary artery disease, but after a year, fewer than 1,600 had been enrolled. Therefore, the original primary outcome of death due to cardiovascular causes or nonfatal myocardial infarction was changed to include coronary revascularization, reducing the sample size to 8,100 [74]. The Carvedilol Post-Infarct Survival

Control in Left Ventricular Dysfunction (CAPRICORN) trial [75] had both poor participant recruitment and lower than expected event rate. To the original primary outcome of all-cause mortality was added a second primary outcome of all-cause mortality or hospitalization for cardiovascular reasons. In order to keep the overall type 1 error at 0.05, the α was divided between the two primary outcomes. Unfortunately, at the end of the trial, there was little difference between groups in the new primary outcome, but a reduction in the original outcome. Had it not been changed, requiring a more extreme result, it would have reached statistical significance [76].

In these examples, the rationale for the change was clearly stated. On occasion, however, the reported primary response variable was changed without clear rationale (or even disclosed in the publication) and after the data had been examined [77, 78]. A survey by Chan et al. [79] found that over 60% of trials conducted in Denmark in 1994–1995 had primary outcome changes between the original protocol and the publication.

Changing the primary outcome during the trial cannot be undertaken lightly and is generally discouraged. It should only be done if other approaches to completing the trial and achieving adequate power are not feasible or affordable. Importantly, it must be done without knowledge of outcome trends. One possible way is for the protocol to specify that if recruitment is below a certain level or overall event rate is under a certain percent, the primary outcome will be changed. Anyone aware of the outcome trends by study group should not be involved in the decision. This includes the data monitoring committee. Sometimes, an independent committee that is kept ignorant of outcome trends is convened to make recommendations regarding the proposed change.

General Comments

Although this text attempts to provide straightforward concepts concerning the selection of study response variables, things are rarely as simple as one would like them to be. Investigators often encounter problems related to design, data monitoring and ethical issues and interpretation of study results.

In long-term studies of participants at high-risk, when total mortality is not the primary response variable, many may nevertheless die. They are, therefore, removed from the population at risk of developing the response variable of interest. Even in relatively short studies, if the participants are seriously ill, death may occur. In designing studies, therefore, if the primary response variable is a continuous measurement, a nonfatal event, or cause-specific mortality, the investigator needs to consider the impact of total mortality for two reasons. First, it will reduce the effective sample size. One might allow for this reduction by estimating the overall mortality and increasing sample size accordingly.

Second, if mortality is related to the intervention, either favorably or unfavorably, excluding from study analysis those who die may bias results for the primary response variable.

One solution, whenever the risk of mortality is high, is to choose total mortality as the primary response variable. Alternatively, the investigator can combine total mortality with a pertinent nonfatal event as a combined primary response variable. Neither of these solutions may be appropriate and, in that case, the investigator should monitor total mortality as well as the primary response variable. Evaluation of the primary response variable will then need to consider those who died during the study, or else the censoring may bias the comparison.

Investigators need to monitor total mortality-as well as any other adverse occurrence-during a study, regardless of whether or not it is the primary response variable (see Chap. 16). The ethics of continuing a study which, despite a favorable trend for the primary response variable, shows equivocal or even negative results for secondary response variables, or the presence of major adverse effects, are questionable. Deciding what to do is difficult if an intervention is giving promising results with regard to death from a specific cause (which may be the primary response variable), yet total mortality is unchanged or increased. An independent data monitoring committee has proved extremely valuable in such circumstances (Chap. 16).

Finally, conclusions from data are not always clear-cut. Issues such as alterations in quality of life or annoying long-term adverse effects may cloud results that are clear with regard to primary response variables such as increased survival. In such circumstances, the investigator must offer her best assessment of the results but should report sufficient detail about the study to permit others to reach their own conclusions (Chap. 21).

References

1. Moyé LA. Random research. *Circulation* 2001;103:3150–3153.
2. Cutler SJ, Greenhouse SW, Cornfield J, Schneiderman MA. The role of hypothesis testing in clinical trials: biometrics seminar. *J Chronic Dis* 1966;19:857–882.
3. Al-Marzouki S, Roberts I, Marshall T, Evans S. The effect of scientific misconduct on the results of clinical trials: a Delphi survey. *Contemp Clin Trials* 2005;26:331–337.
4. Angell M. Caring for women's health - what is the problem? (editorial). *N Engl J Med* 1993;329:271–272.
5. NIH Revitalization Act, Subtitle B, Part 1, Sec. 131–133, June 10, 1993.
6. Simes RJ, O'Connell RL, Aylward PE, et al. Unexplained international differences in clinical outcomes after acute myocardial infarction and fibrinolytic therapy: lessons from the Hirulog and Early Reperfusion or Occlusion (HERO)-2 trial. *Am Heart J* 2010;159:988–997.
7. Cannon CP, Harrington RA, James S, et al. Comparison of ticagrelor with clopidogrel in patients with a planned invasive strategy for acute coronary syndromes (PLATO): a randomised double-blind study. *Lancet* 2010;375:283–293.

8. Mahaffey KW, Wojdyla DM, Carroll K, et al. Ticagrelor compared with clopidogrel by geographic region in the Platelet Inhibition and Patient Outcomes (PLATO) trial. *Circulation* 2011;124:544–554.
9. Freedman LS, Simon R, Foulkes MA, et al. Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993 - the perspective of NIH clinical trialists. *Control Clin Trials* 1995;16:277–285.
10. Piantadosi S, Wittes J. Letter to the editor. *Control Clin Trials* 1993;14:562–567.
11. ISIS-2 (Second International Study of Infarct Survival) 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; *ii*:349–360.
12. Shimkin MB. The problem of experimentation on human beings. I. The research worker's point of view. *Science* 1953;117:205–207.
13. Chalmers TC. Invited Remarks: National Conference on Clinical Trials Methodology. *Clin Pharmacol Ther* 1979;25:649–650.
14. Stamler J. Invited Remarks: National Conference on Clinical Trials Methodology. *Clin Pharmacol Ther* 1979;25:651–654.
15. The Cardiac Arrhythmia Suppression Trial II Investigators. Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *N Engl J Med* 1992;327:227–233.
16. Miller AB, Buring J, Williams OD. Stopping the carotene and retinol efficacy trial: the viewpoint of the safety and endpoint monitoring committee. In DeMets DL, Furberg CD, Friedman LM.(eds). *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York: Springer 2006, pp.220-227.
17. Swedberg K, Held P, Kjekshus J, et al. Effects of the early administration of enalapril on mortality in patients with acute myocardial infarction – results of the Cooperative New Scandinavian Enalapril Survival Study II (Consensus II). *N Engl J Med* 1992;327:678–684.
18. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med* 1993;329:673–682.
19. Lee KL, Woodlief LH, Topol EJ, et al. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction. Results from an international trial of 41,021 patients. GUSTO-1 Investigators. *Circulation* 1995;91:1659–1668.
20. APEX AMI Investigators, Armstrong PW, Granger CB, Adams PX, et al. Pexelizumab for acute ST-elevation myocardial infarction in patients undergoing primary percutaneous coronary intervention: a randomized controlled trial. *JAMA* 2007;297:43–51.
21. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N Engl J Med* 1991;325: 293–302
22. Ross AM, Coyne KS, Moreyra E, et al. Extended mortality benefit of early postinfarction reperfusion. GUSTO-1 Angiographic Investigators. Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries Trial. *Circulation* 1998;97:1549–1556.
23. Simes RJ, Topol EJ, Holmes DR Jr., et al. Link between the angiographic substudy and mortality outcomes in a large randomized trial of myocardial reperfusion. Importance of early and complete infarct artery reperfusion. GUSTO-1 Investigators. *Circulation* 1995;91:1923–1928.
24. Stone GW, Witzenbichler B, Guagliumi G, et al. Bivalirudin during primary PCI in acute myocardial infarction. *N Engl J Med* 2008;358:2218–2230.
25. Stone GW, Clayton T, Deliargyris EN, Prats J, Mehran R, Pocock SJ. Reduction in cardiac mortality with bivalirudin in patients with and without major bleeding: the HORIZONS-AMI trial (Harmonizing Outcomes with Revascularization and Stents in Acute Myocardial Infarction). *J Am Coll Cardiol* 2014;63:15–20.
26. Mega JL, Close SL, Wiviott SD, et al. Cytochrome P-450 polymorphisms and response to clopidogrel. *N Engl J Med* 2009;360:354–362.
27. Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC Jr. Scientific evidence underlying the ACC/AHA clinical practice guidelines. *JAMA* 2009;301:831–841.

28. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409–422.
29. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003;290:1624–1632.
30. Eapen ZN, Lauer MS, Temple RJ. The imperative of overcoming barriers to the conduct of large, simple trials. *JAMA* 2014;311:1397–1398.
31. Palmerini T, Biondi-Zoccai G, Della Riva D, et al. Clinical outcomes with drug-eluting and bare-metal stents in patients with ST-segment elevation myocardial infarction: evidence from a comprehensive network meta-analysis. *J Am Coll Cardiol* 2013;62:496–504.
32. McFadden E, Stabile E, Regar E, et al. Late thrombosis in drug-eluting coronary stents after discontinuation of antiplatelet therapy. *Lancet* 2004;364:1519–1521.
33. Ong AT, McFadden EP, Regar E, et al. Late angiographic stent thrombosis (LAST) events with drug eluting stents. *J Am Coll Cardiol* 2005;45:2088–2092.
34. Mauri L, Hsieh W-H, Massaro JM, et al. Stent thrombosis in randomized clinical trials of drug-eluting stents. *N Engl J Med* 2007;356:1020–1029.
35. Metal-on-Metal Hip Implants: FDA Safety Communication. http://www.fda.gov/medicaldevices/safety/alertsandnotices/ucm335775.htm
36. Heneghan C, Langton D, Thompson M. Ongoing problems with metal-on-metal hip implants. BMJ 2012;344:e1349. http://www.bmj.com/content/344/bmj.e1349
37. Urokinase Pulmonary Embolism Trial Study Group. Urokinase Pulmonary Embolism Trial: phase I results. *JAMA* 1970;214:2163–2172.
38. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ* 2010;341:c3920.
39. Tomlinson G, Detsky AS. Composite end points in randomized trials: there is no free lunch. *JAMA* 2010;303:267–268.
40. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 1989;321:129–135.
41. Cairns J, Cohen L, Colton T, et al. Data Monitoring Board of the Physicians' Health Study. Issues in the early termination of the aspirin component of the Physicians' Health Study. *Ann Epidemiol* 1991;1:395–405.
42. Schwartz GG, Olsson AG, Ezekowitz MD, et al for the Myocardial Ischemia Reduction with Aggressive Cholesterol Lowering (MIRACL) Study Investigators. Effects of atorvastatin on early ischemic events in acute coronary syndromes: the MIRACL study: a randomized controlled trial. *JAMA* 2001;285:1711–1718.
43. 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. *Br Med J* 2007; 334:756–757.
44. Lim E, Brown A, Helmy A, et al. Composite outcomes in cardiovascular research: a survey of randomized trials. *Ann Intern Med* 2008;149:612–617.
45. Neaton JD, Wentworth DN, Rhame F, et al. Methods of studying intervention: considerations in choice of a clinical endpoint for AIDS clinical trials. *Stat Med* 1994;13:2107–2125.
46. Hallstrom AP, Litivin PE, Weaver WD. A method of assigning scores to the components of a composite outcome: an example from the MITI trial. *Control Clinl Trials* 1992;13:148–155.
47. Mentzer RM Jr, Bartels C, Bolli R, et al. Sodium-hydrogen exchange inhibition by cariporide to reduce the risk of ischemic cardiac events in patients undergoing coronary artery bypass grafting: results of the EXPEDITION study. *Ann Thorac Surg* 2008;85:1261–1270.
48. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med* 1989; 8:415–425.
49. Fleming T, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? *Ann Intern Med* 1996;125:605–613.
50. Fleming TR. Surrogate markers in AIDS and cancer trials. *Stat Med* 1994;13:1423–1435.

51. Bigger JT Jr, Fleiss JL, Kleiger R, et al, and the Multicenter Post-Infarction Research Group. The relationships among ventricular arrhythmias, left ventricular dysfunction, and mortality in the 2 years after myocardial infarction. *Circulation* 1984;69:250–258.
52. Vlay SC. How the university cardiologist treats ventricular premature beats: a nationwide survey of 65 university medical centers. *Am Heart J* 1985;110:904–912
53. Morganroth J, Bigger JT Jr, Anderson JL. Treatment of ventricular arrhythmias by United States cardiologists: a survey before the Cardiac Arrhythmia Suppression Trial (CAST) results were available. *Am J Cardiol* 1990;65:40–48.
54. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary Report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;321:406–412.
55. Packer M. Vasodilator and inotropic drugs for the treatment of chronic heart failure: distinguishing hype from hope. *J Am Coll Cardiol*. 1988;12:1299–1317.
56. Packer M, Carver JR, Rodehoffer RT, et al. Effect of oral milrinone on mortality in severe chronic heart failure. *N Engl J Med* 1991;325:1468–1475.
57. The Xamoterol in Severe Heart Failure Study Group. Xamoterol in severe heart failure. *Lancet* 1990;336:1–6; correction *Lancet* 1990;336:698.
58. Barter PJ, Caulfield M, Eriksson M, et al for the ILLUMINATE Investigators. Effects of torcetrapib in patients at high risk for coronary events. *N Engl J Med* 2007;357:2109–2122.
59. Boden WE, Sidhu MS, Toth PP. The therapeutic role of niacin in dyslipidemia management. *J Cardiovasc Pharmacol Ther* 2014;19:141–158.
60. AIM-HIGH Investigators, Boden WE, Probstfield JL, Anderson T, et al. Niacin in patients with low HDL cholesterol levels receiving intensive statin therapy. *N Engl J Med* 2011;365:2255–2267.
61. HPS2-THRIVE Collaborative Group. Landray MJ, Haynes R, Hopewell JC, et al. Effects of extended-release niacin with laropiprant in high-risk patients. *N Engl J Med* 2014;371:203–212.
62. Cohen J. Searching for markers on the AIDS trail. *Science* 1992;258:388–390.
63. Lin DY, Fischl MA, Schoenfeld DA. Evaluating the role of CD-4 lymphocyte counts as surrogate endpoints in human immunodeficiency virus clinical trials. *Stat Med* 1993;12:835–842.
64. Choi S, Lagakos SW, Schooley RT, Volberding PA. CD4+ lymphocytes are an incomplete surrogate marker for clinical progression in persons with asymptomatic HIV infection taking zidovidine. *Ann Intern Med* 1993;118:674–680.
65. Fischl MA., Olson RM, Follansbee SE, et al. Zalcitabine compared with zidovudine in patients with advanced HIV-1 infection who received previous zidovudine therapy. *Ann Intern Med* 1993;118:762–769.
66. Aboulker JP, Swart AM. Preliminary analysis of the Concorde trial. *Lancet* 1993;341:889–890.
67. Advanced Colorectal Cancer Meta-Analysis Project. Modulation of fluoruracil by leucovorin in patients with advanced colorectal cancer. Evidence in terms of response rate. *J Clin Oncol* 1992;10:896–903.
68. Riggs BL, Hodgson SF, O'Fallon WM, et al. Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis. *N Engl J Med* 1990;322:802–809.
69. Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Stat Med* 1989;8:431–440.
70. Ray WA, Griffin MR, Avorn J. Sounding Board: Evaluating drugs after their approval for clinical use. *N Engl J Med* 1993;329:2029–2032.
71. FDA Guidance for Industry. Diabetes Mellitus—Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes. http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071627.pdf
72. Avorn J. Viewpoint: Approval of a tuberculosis drug based on a paradoxical surrogate measure. *JAMA* 2013;309:1349–1350.

73. 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.
74. The PEACE Trial Investigators. Angiotensin-converting-enzyme inhibition in stable coronary artery disease. *N Engl J Med* 2004;351:2058–2068.
75. Dargie HJ. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. *Lancet* 2001;357:1385–1390.
76. Julian D. The data monitoring experience in the Carvedilol Post-Infarct Survival Control in Left Ventricular Dysfunction Study: Hazards of changing primary outcomes. In DeMets DL, Furberg CD, Friedman LM (eds.). *Data Monitoring in Clinical Trials*. New York: Springer, 2006, pp. 337–345.
77. The Anturane Reinfarction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *N Engl J Med* 1980;302:250–256.
78. Anturane Reinfarction Trial Policy Committee. The Anturane Reinfarction Trial: reevaluation of outcome. *N Engl J Med* 1982;306:1005–1008.
79. Chan A-W, Hróbjartsson A, Haahr MT, et al. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* 2004;291:2457–2465.