Chapter 5
Basic Study Design

The foundations for the design of controlled experiments were established for agricultural application. They are described in several classical statistics textbooks [1–4]. From these sources evolved the basic design of controlled clinical trials.

Although the history of clinical experimentation contains several instances in which the need for control groups has been recognized [5, 6], this need was not widely accepted until the 1950s [7]. In the past, when a new intervention was first investigated, it was likely to be given to only a small number of people, and the outcome compared, if at all, to that in people with the same condition previously treated in a different manner. The comparison was informal and frequently based on memory alone. Sometimes, in one kind of what has been called a "quasi-experimental" study, people were evaluated initially and then reexamined after an intervention had been introduced. In such studies, the changes from the initial state were used as the measure of success or failure of the new intervention. What could not be known was whether the person would have responded in the same manner if there had been no intervention at all. However, then—and sometimes even today—this kind of observation has formed the basis for the use of new interventions.

Of course, some results are so highly dramatic that no comparison group is needed. Successful results of this magnitude, however, are rare. One example is the effectiveness of penicillin in pneumococcal pneumonia. Another example originated with Pasteur who in 1884 was able to demonstrate that a series of vaccine injections protected dogs from rabies [8]. He suggested that due to the long incubation time, prompt vaccination of a human being after infection might prevent the fatal disease. The first patient was a 9-year-old boy who had been bitten 3 days earlier by a rabid dog. The treatment was completely effective. Confirmation came from another boy who was treated within 6 days of having been bitten. During the next few years, hundreds of patients were given the anti-rabies vaccine. If given within certain time-limits, it was almost always effective.

Gocke reported on a similar, uncontrolled study of patients with acute fulminant viral hepatitis [9]. Nine consecutive cases had been observed, all of whom had a fatal outcome. The next diagnosed case, a young staff nurse in hepatic coma, was

L.M. Friedman et al., *Fundamentals of Clinical Trials*,
DOI 10.1007/978-3-319-18539-2_5

given immunotherapy in addition to standard treatment. The patient survived as did four others among eight given the antiserum. The author initially thought that this uncontrolled study was conclusive. However, in considering other explanations for the encouraging findings, he could not eliminate the possibility that a tendency to treat patients earlier in the course and more intensive care might be responsible for the observed outcome. Thus, he joined a double-blind, randomized trial comparing hyperimmune anti-Australia globulin to normal human serum globulin in patients with severe acute hepatitis. Nineteen of 28 patients (67.9%) randomized to control treatment died, compared to 16 of 25 patients (64%) randomized to treatment with exogenous antibody, a statistically nonsignificant difference [10].

A number of medical conditions are either of short duration or episodic in nature. Evaluation of therapy in these cases can be difficult in the absence of controlled studies. Snow and Kimmelman reviewed various uncontrolled studies of surgical procedures for Ménière's disease [11]. They found that about 75% of patients improved, but noted that this is similar to the 70% remission rate occurring without treatment.

Given the wide spectrum of the natural history of almost any disease and the variability of an individual's response to an intervention, most investigators recognize the need for a defined control or comparison group.

Fundamental Point

Sound scientific clinical investigation almost always demands that a control group be used against which the new intervention can be compared. Randomization is the preferred way of assigning participants to control and intervention groups.

Overview

Statistics and epidemiology textbooks and papers [12–31], cover various study designs in some detail. Green and Byar also present a "hierarchy of strength of evidence concerning efficacy of treatment" [32]. In their scheme, anecdotal case reports are weakest and confirmed randomized clinical trials are strongest, with various observational and retrospective designs in between. This chapter will discuss several major clinical trial designs.

Most trials use the so-called parallel design. That is, the intervention and control groups are followed simultaneously from the time of allocation to one or the other. Exceptions to the simultaneous follow-up are historical control studies. These compare a group of participants on a new intervention with a previous group of participants on standard or control therapy. A modification of the parallel design is the cross-over trial, which uses each participant at least twice, at least once as a member of the control group and at least once as a member of one or more

intervention groups. Another modification is a withdrawal study, which starts with all participants on the active intervention and then, usually randomly, assigns a portion to be followed on the active intervention and the remainder to be followed off the intervention. Factorial design trials, as described later in this chapter, employ two or more independent assignments to intervention or control.

Regardless of whether the trial is a typical parallel design or some variant, one must select the kind of control group and the way participants are allocated to intervention or control. Controls may be on placebo, no treatment, usual or standard care, or a specified treatment. Randomized control and nonrandomized concurrent control studies both assign participants to either the intervention or the control group, but only the former makes the assignment by using a random procedure. Hybrid designs may use a combination of randomized and non-randomized controls. Large, simple trials or pragmatic trials generally have broader and simpler eligibility criteria than other kinds of trials, but as with other studies, can use any of the indicated controls. Allocation to intervention or control may also be done differently, even if randomized. Randomization may be by individual participant or by groups of participants (group or cluster assignment). Adaptive designs may adjust intervention or control assignment or sample size on the basis of participant characteristics or outcomes.

Finally, there are superiority trials and equivalence or noninferiority trials. A *superiority trial*, which for many years was the typical kind of trial, assesses whether the new intervention is different from (better or worse than) the control. An *equivalence trial* would assess if the new intervention is more or less equal to the control. A *noninferiority trial* evaluates whether the new intervention is no worse than the control by some margin, delta (δ). In both of these latter cases, the control group would be on a treatment that had previously been shown to be effective, i.e., have an active control.

Questions have been raised concerning the method of selection of the control group, but the major controversy in the past revolved around the use of historical versus randomized control [33–35]. With regard to drug evaluation, this controversy is less intense than in the past. It has been hotly contested, however, in the evaluation of new devices or procedures [36, 37]. While it is acknowledged that randomized controls provide the best evidence, devices that are relatively little used may be approved based on historical controls with post-marketing studies to further assess possible adverse effects. An example is a device used for closure of a cardiac chamber wall defect [38]. It should be noted that after marketing, rare, but serious adverse effects were reported [39]. No study design is perfect or can answer all questions. Each of the designs has advantages and disadvantages, but a randomized control design is the standard by which other studies should be judged. A discussion of sequential designs is postponed until Chap. 17 because the basic feature involves interim analyses.

For each of the designs it is assumed, for simplicity of discussion, that a single control group and a single intervention group are being considered. These designs can be extended to more than one intervention group and more than one control group.

Randomized Control Trials

Randomized control trials are comparative studies with an intervention group and a control group; the assignment of the participant to a group is determined by the formal procedure of randomization. Randomization, in the simplest case, is a process by which all participants are equally likely to be assigned to either the intervention group or the control group. The features of this technique are discussed in Chap. 6. There are three advantages of the randomized design over other methods for selecting controls [35].

First, randomization removes the potential of bias in the allocation of participants to the intervention group or to the control group. Such selection bias could easily occur, and cannot be necessarily prevented, in the non-randomized concurrent or historical control study because the investigator or the participant may influence the choice of intervention. This influence can be conscious or subconscious and can be due to numerous factors, including the prognosis of the participant. The direction of the allocation bias may go either way and can easily invalidate the comparison. This advantage of randomization assumes that the procedure is performed in a valid manner and that the assignment cannot be predicted (see Chap. 6).

Second, somewhat related to the first, is that randomization tends to produce comparable groups; that is, measured as well as unknown or unmeasured prognostic factors and other characteristics of the participants at the time of randomization will be, on the average, evenly balanced between the intervention and control groups. This does not mean that in any single experiment all such characteristics, sometimes called baseline variables or covariates, will be perfectly balanced between the two groups. However, it does mean that for independent covariates, whatever the detected or undetected differences that exist between the groups, the overall magnitude and direction of the differences will tend to be equally divided between the two groups. Of course, many covariates are strongly associated; thus, any imbalance in one would tend to produce imbalances in the others. As discussed in Chaps. 6 and 18, stratified randomization and stratified analysis are methods commonly used to guard against and adjust for imbalanced randomizations (i.e., "accidental" bias).

Third, the validity of statistical tests of significance is guaranteed. As has been stated [35], "although groups compared are never perfectly balanced for important covariates in any single experiment, the process of randomization makes it possible to ascribe a probability distribution to the difference in outcome between treatment groups receiving equally effective treatments and thus to assign significance levels to observed differences." The validity of the statistical tests of significance is not dependent on the balance of the prognostic factors between the randomized groups. The chi-square test for two-by-two tables and Student's *t*-test for comparing two means can be justified on the basis of randomization alone without making further assumptions concerning the distribution of baseline variables. If randomization is not used, further assumptions concerning the comparability of the groups and the appropriateness of the statistical models must be made before the comparisons will be valid. Establishing the validity of these assumptions may be difficult.

In 1977, randomized and nonrandomized trials of the use of anticoagulant therapy in patients with acute myocardial infarctions were reviewed by Chalmers et al. and the conclusions compared [40]. Of 32 studies, 18 used historical controls and involved a total of 900 patients, 8 used nonrandomized concurrent controls and involved over 3,000 patients, and 6 were randomized trials with a total of over 3,800 patients. The authors reported that 15 of the 18 historical control trials and 5 of the 8 nonrandomized concurrent control trials showed statistically significant results favoring the anticoagulation therapy. Only one of the six randomized control trials showed significant results in support of this therapy. Pooling the results of these six randomized trials yielded a statistically significant 20% reduction in total mortality, confirming the findings of the nonrandomized studies. Pooling the results of the nonrandomized control studies showed a reduction of about 50% in total mortality in the intervention groups, more than twice the decrease seen in the randomized trials. Peto [41] has assumed that this difference in reduction is due to bias. He suggests that since the presumed bias in the nonrandomized trials was of the same order of magnitude as the presumed true effect, the non-randomized trials could have yielded positive answers even if the therapy had been of no benefit. Of course, pooling results of several studies can be hazardous. As pointed out by Goldman and Feinstein [42], not all randomized trials of anticoagulants study the same kind of participants, use precisely the same intervention or measure the same response variables. And, of course, not all randomized trials are done equally well. The principles of pooled analysis, or meta-analysis, are covered in Chap. 18.

In the 1960s, Grace, Muench and Chalmers [43] reviewed studies involving portacaval shunt operations for patients with portal hypertension from cirrhosis. In their review, 34 of 47 non-randomized studies strongly supported the shunt procedure, while only one of the four randomized control trials indicated support for the operation. The authors concluded that the operation should not be endorsed.

Sacks and coworkers expanded the work by Chalmers et al. referenced above [40], to five other interventions [44]. They concluded that selection biases led historical control studies to favor inappropriately the new interventions. It was also noted that many randomized control trials were of inadequate size, and therefore may have failed to find benefits that truly existed [45]. Chalmers and his colleagues also examined 145 reports of studies of treatment after myocardial infarction [46]. Of the 57 studies that used a randomization process that had proper concealment of allocation to intervention or control, 14% had at least one significant ($p < 0.05$) maldistribution of baseline variables with 3.4% of all of the variables significantly different between treatment groups. Of these 57 studies, 9% found significant outcome differences between groups. Among the 43 reports where the control groups were selected by means of a nonrandom process, 58% had baseline variable differences and 34% of all of the variables were significantly different between groups. The outcomes between groups in the nonrandom studies were significantly different 58% of the time. For the 45 studies that used a randomized, but unblinded process to select the control groups, the results were in between; 28% had baseline imbalances, 7% of the baseline variables were significantly different, and 24% showed significant outcome differences.

The most frequent objections to the use of the randomized control clinical trial were stated by Ingelfinger [47], to be "emotional and ethical." Many clinicians feel that they must not deprive a participant from receiving a new therapy or intervention which they, or someone else, believe to be beneficial, regardless of the validity of the evidence for that claim. The argument aimed at randomization is that in the typical trial it deprives about one-half the participants from receiving the new and presumed better intervention. There is a large literature on the ethical aspects of randomization. See Chap. 2 for a discussion of this issue.

Not all clinical studies can use randomized controls. Occasionally, the prevalence of the disease is so rare that a large enough population can not be readily obtained. In such an instance, only case-control studies might be possible. Such studies, which are not clinical trials according to the definition in this book, are discussed in standard epidemiology textbooks [15, 16, 22, 28].

Zelen proposed a modification of the standard randomized control study [48]. He argued that investigators are often reluctant to recruit prospective trial participants not knowing to which group the participant will be assigned. Expressing ignorance of optimal therapy compromises the traditional doctor-patient relationship. Zelen, therefore, suggested randomizing eligible participants before informing them about the trial. Only those assigned to active intervention would be asked if they wish to participate. The control participants would simply be followed and their outcomes monitored. Obviously, such a design could not be blinded. Another major criticism of this controversial design centers around the ethical concern of not informing participants that they are enrolled in a trial. The efficiency of the design has also been evaluated [49]. It depends on the proportion of participants consenting to comply with the assigned intervention. To compensate for this possible inefficiency, one needs to increase the sample size (Chap. 8). The Zelen approach has been tried with varying degrees of success [50, 51]. Despite having been proposed in 1979 it does not appear to have been widely used.

Nonrandomized Concurrent Control Studies

Controls in this type of study are participants treated without the new intervention at approximately the same time as the intervention group is treated. Participants are allocated to one of the two groups, but by definition this is not a random process. An example of a nonrandomized concurrent control study would be a comparison of survival results of patients treated at two institutions, one institution using a new surgical procedure and the other using more traditional medical care. Another example is when patients are offered either of two treatments and the patient selects the one that he or she thinks is preferable. Comparisons between the two groups is then made, adjusting for any observed baseline imbalances.

To some investigators, the nonrandomized concurrent control design has advantages over the randomized control design. Those who object to the idea of ceding to chance the responsibility for selecting a person's treatment may favor this design.

It is also difficult for some investigators to convince potential participants of the need for randomization. They find it easier to offer the intervention to some and the control to others, hoping to match on key characteristics.

The major weakness of the nonrandomized concurrent control study is the potential that the intervention group and control group are not strictly comparable. It is difficult to prove comparability because the investigator must assume that she has information on all the important prognostic factors. Selecting a control group by matching on more than a few factors is impractical and the comparability of a variety of other characteristics would still need to be evaluated. In small studies, an investigator is unlikely to find real differences which may exist between groups before the initiation of intervention since there is poor sensitivity statistically to detect such differences. Even for large studies that could detect most differences of real clinical importance, the uncertainty about the unknown or unmeasured factors is still of concern.

Is there, for example, some unknown and unmeasurable process that results in one type of participant's being recruited more often into one group and not into the other? If all participants come from one institution, physicians may select participants into one group based on subtle and intangible factors. In addition, there exists the possibility for subconscious bias in the allocation of participants to either the intervention or control group. One group might come from a different socioeconomic class than the other group. All of these uncertainties will decrease the credibility of the concurrent but nonrandomized control study. For any particular question, the advantages of reduced cost, relative simplicity and investigator and participant acceptance must be carefully weighed against the potential biases before a decision is made to use a non-randomized concurrent control study. We believe this will occur very rarely.

Historical Controls and Databases

In historical control studies, a new intervention is used in a series of participants and the results are compared to the outcome in a previous series of comparable participants. Historical controls are thus, by this definition, nonrandomized and nonconcurrent.

Strengths of Historical Control Studies

The argument for using a historical control design is that all new participants can receive the new intervention. As presented by Gehan and Freireich [33] many clinicians believe that no participant should be deprived of the possibility of receiving a new therapy or intervention. Some clinicians require less supportive evidence than others to accept a new intervention as being beneficial. If an investigator is already of the opinion that the new intervention is beneficial, then

she would most likely consider any restriction on its use unethical. Therefore, she would favor a historical control study. In addition, participants may be more willing to enroll in a study if they can be assured of receiving a particular therapy or intervention. Finally, since all new participants will be on the new intervention, the time required to complete recruitment of participants for the trial will be cut approximately in half. This allows investigators to obtain results faster or do more studies with given resources. Alternatively, the sample size for the intervention group can be larger, with increased power.

Gehan emphasized the ethical advantages of historical control studies and pointed out that they have contributed to medical knowledge [52]. Lasagna argued that medical practitioners traditionally relied on historical controls when making therapeutic judgments. He maintained that, while sometimes faulty, these judgments are often correct and useful [53].

Typically, historical control data can be obtained from two sources. First, control group data may be available in the literature. These data are often undesirable because it is difficult, and perhaps impossible, to establish whether the control and intervention groups are comparable in key characteristics at the onset. Even if such characteristics were measured in the same way, the information may not be published and for all practical purposes it will be lost. Second, data may not have been published but may be available on computer files or in medical charts. Such data on control participants, for example, might be found in a large center which has several ongoing clinical investigations. When one study is finished, the participants in that study may be used as a control group for some future study. Centers which do successive studies, as in cancer research, will usually have a system for storing and retrieving the data from past studies for use at some future time. The advent of electronic medical records may also facilitate access to historical data from multiple sources, although it does not solve the problem of nonstandard and variable assessment or missing information.

Limitations of Historical Control Studies

Despite the time and cost benefits, as well as the ethical considerations, historical control studies have potential limitations which should be kept in mind. They are particularly vulnerable to bias. Moertel [54] cited a number of examples of treatments for cancer which have been claimed, on the basis of historical control studies, to be beneficial. Many treatments in the past were declared breakthroughs on the basis of control data as old as 30 years. Pocock [55] identified 19 instances of the same intervention having been used in two consecutive trials employing similar participants at the same institution. Theoretically, the mortality in the two groups using the same treatment should be similar. Pocock noted that the difference in mortality rates between such groups ranged from negative 46% to plus 24%. Four of the 19 comparisons of the same intervention showed differences significant at the 5% level.

An improvement in outcome for a given disease may be attributed to a new intervention when, in fact, the improvement may stem from a change in the patient

population or patient management. Shifts in patient population can be subtle and perhaps undetectable. In a Veterans Administration Urological Research Group study of prostate cancer [56], 2,313 people were randomized to placebo or estrogen treatment groups over a 7-year period. For those enrolled during the last 2–3 years, no differences were found between the placebo and estrogen groups. However, those assigned to placebo entering in the first 2–3 years had a shorter survival time than those assigned to estrogen entering in the last 2–3 years of the study. The reason for the early apparent difference is probably that the people randomized earlier were older than the later group and thus were at higher risk of death during the period of observation [35]. The results would have been misleading had this been a historical control study and had a concurrent randomized comparison group not been available.

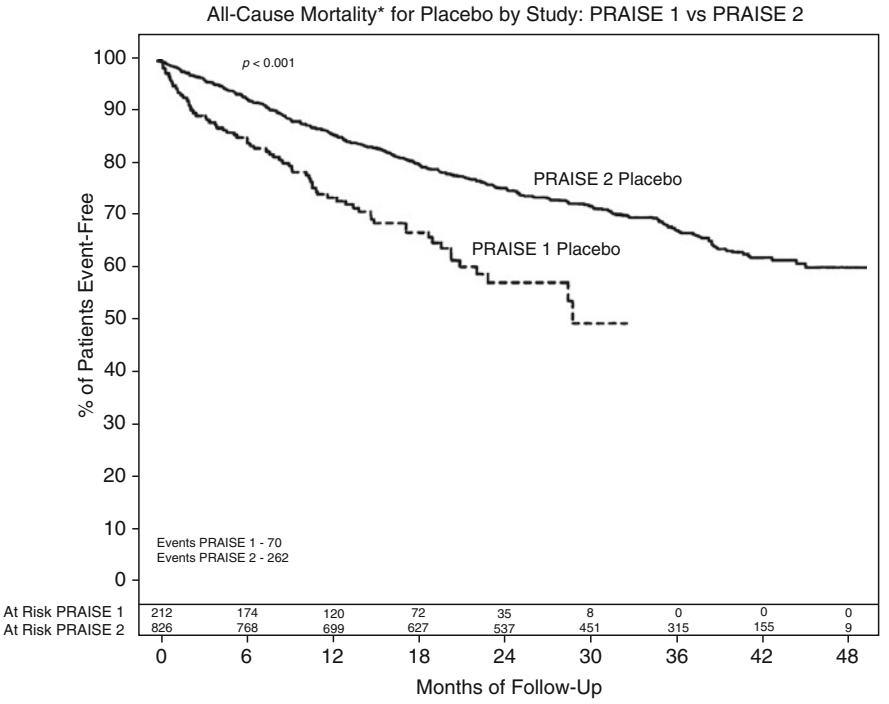
A more recent example involves two trials evaluating the potential benefit of amlodipine, a calcium channel blocker, in patients with heart failure. The first trial, the Prospective Randomized Amlodipine Survival Evaluation trials, referred to as PRAISE-I [57], randomized participants to amlodipine or placebo, stratifying by ischemic or nonischemic etiology of the heart failure. The primary outcome, death plus hospitalization for cardiovascular reasons, was not significantly different between groups ($p = 0.31$), but the reduction in mortality almost reached significance ($p = 0.07$). An interaction with etiology was noted, with all of the benefit from amlodipine in both the primary outcome and mortality seen in those with nonischemic etiology. A second trial, PRAISE-2 [58], was conducted in only those with nonischemic causes of heart failure. The impressive subgroup findings noted in PRAISE-1 were not replicated. Of relevance here is that the event rates in the placebo group in PRAISE-2 were significantly lower than in the nonischemic placebo participants from the first trial (see Fig. 5.1).

Even though the same investigators conducted both trials using the same protocol, the kinds of people who were enrolled into the second trial were markedly different from the first trial. Covariate analyses were unable to account for the difference in outcome.

On a broader scale, for both known and unknown reasons, in many countries trends in prevalence of various diseases occur [59]. Therefore, any clinical trial in those conditions, involving long-term therapy using historical controls would need to separate the treatment effect from the time trends, an almost impossible task. Examples are seen in Figs. 5.2 and 5.3.

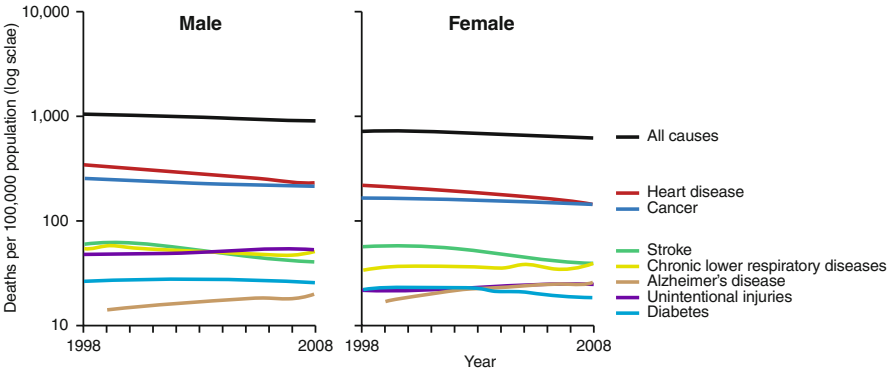
Figure 5.2 illustrates the changes over time, in rates of the leading causes of death in the United States [60]. A few of the causes exhibit quite large changes. Figure 5.3 shows incidence of hepatitis in the U.S. [61]. The big changes make interpretation of historical control trials difficult.

The method by which participants are selected for a particular study can have a large impact on their comparability with earlier participant groups or general population statistics. In the Coronary Drug Project [62], a trial of survivors of myocardial infarction initiated in the 1960s, an annual total mortality rate of 6% was anticipated in the control group based on rates from a fairly unselected group of myocardial infarction patients. In fact, a control group mortality rate of about 4% was observed, and no significant differences in mortality were seen between the intervention groups and the control group. Using the historical control approach, a



Information for PRAISE 2 is from the ENDPT dataset sent to SDAC on December 19, 1999. The PRAISE 1 results are for the non-ischemic subgroup only.*For PRAISE 1, transplants have been censored at the time of transplant and are not considered an event for this analysis. For PRAISE 2, patients with transplants are followed for survival post-transplant.

Fig. 5.1 PRAISE 1 and 2 placebo arms



NOTE: Death rates are age-adjusted. Starting with 1999 data, cause of death is coded according to ICD-10
SOURCE: CDC/NCHS, Health United States, 2011 [60]. Data from the National Vital Statistics System

Fig. 5.2 Death rates for selected causes of death for all ages, by sex: United States, 1998–2008

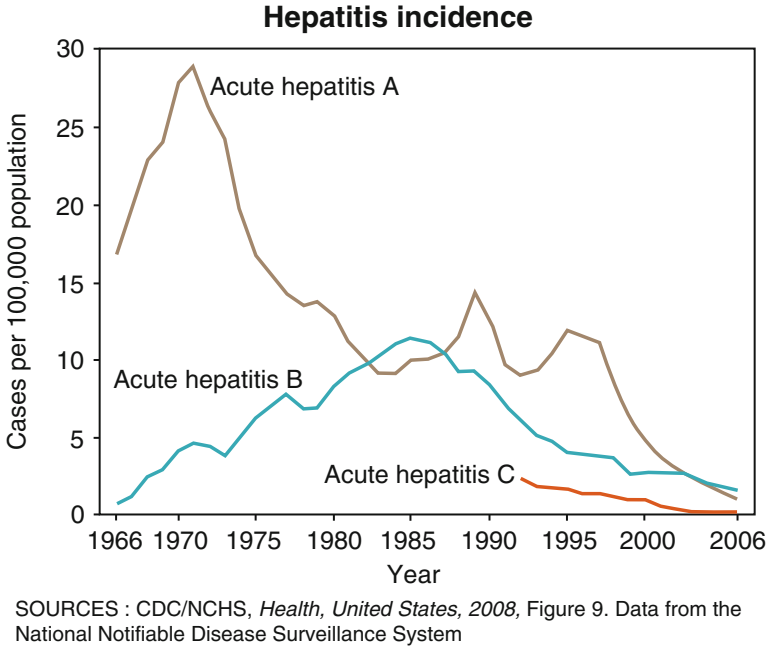


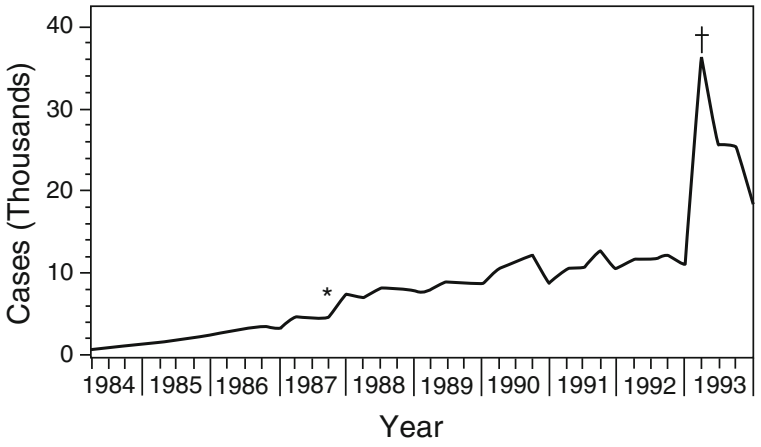
Fig. 5.3 Changes in incidence of hepatitis, by type, in the U.S. [61]

33% reduction in mortality might have been claimed for the treatments. One explanation for the discrepancy between anticipated and observed mortality is that entry criteria excluded those most seriously ill.

Shifts in diagnostic criteria for a given disease due to improved technology can cause major changes in the recorded frequency of the disease and in the perceived prognosis of those with the disease. The use of elevated serum troponin, sometimes to the exclusion of the need for other features of an acute myocardial infarction such as symptoms or electrocardiographic changes, has clearly led to the ability to diagnose more infarctions. Changes in the kinds of troponin measured and in how it is used to define myocardial infarction can also affect reported incidence. Conversely, the ability to abort an evolving infarction by means of percutaneous coronary intervention or thrombolytic therapy, can reduce the number of clearly diagnosed infarctions.

In 1993, the Centers for Disease Control and Prevention (CDC) in the U.S. implemented a revised classification system for HIV infection and an expanded surveillance case definition of AIDS. This affected the number of cases reported [63, 64]. See Fig. 5.4.

International coding systems and names of diseases change periodically and, unless one is aware of the modifications, prevalence of certain conditions can appear to change abruptly. For example, when the Eighth Revision of the International Classification of Diseases came out in 1968, almost 15% more deaths were assigned to ischemic heart disease than had been assigned in the Seventh Revision [65]. When the Ninth Revision



* Case definition revised in October 1987 to include additional illnesses and to revise diagnostic criteria (3).
† Case definition revised in 1993 to include CD4+ criteria and three illnesses (pulmonary tuberculosis, recurrent pneumonia, and invasive cervical cancer) (1).

Fig. 5.4 AIDS cases, by quarter year of report—United States, 1984–1993 [64]

appeared in 1979, there was a correction downward of a similar magnitude [66]. The transition to the Tenth Revision will also lead to changes in assignment of causes of deaths [67]. A common concern about historical control designs is the accuracy and completeness with which control group data are collected. With the possible exception of special centers which have many ongoing studies, data are generally collected in a nonuniform manner by numerous people with diverse interests in the information. Lack of uniform collection methods can easily lead to incomplete and erroneous records. Data on some important prognostic factors may not have been collected at all. Because of the limitations of data collected historically from medical charts, records from a center which conducts several studies and has a computerized data management system may provide the most reliable historical control data.

Role of Historical Controls

Despite the limitations of the historical control study, it does have a place in scientific investigation. As a rapid, relatively inexpensive method of obtaining initial impressions regarding a new therapy, such studies can be important. This is particularly so if investigators understand the potential biases and are willing to miss effective new therapies if bias works in the wrong direction. Bailar et al. [68] identified several features which can strengthen the conclusions to be drawn from historical control studies. These include an a priori identification of a reasonable hypothesis and planning for analysis.

In some special cases where the diagnosis of a disease is clearly established and the prognosis is well known or the disease highly fatal, a historical control study may be the only reasonable design. The results of penicillin in treatment of pneumococcal pneumonia were so dramatic in contrast to previous experience that no further evidence was really required. Similarly, the benefits of treatment of malignant hypertension became readily apparent from comparisons with previous, untreated populations [69–71].

The use of prospective registries to characterize patients and evaluate effects of therapy has been advocated [72–74]. Supporters say that a systematic approach to data collection and follow-up can provide information about the local patient population, and can aid in clinical decision making. They argue that clinical trial populations may not be representative of the patients actually seen by a physician. Moon et al. described the use of databases derived from clinical trials to evaluate therapy [75]. They stress that the high quality data obtained through these sources can reduce the limitations of the typical historical control study. Many hospitals and other large medical care systems have electronic health records. Other clinical care entities are more slowly converting to electronic systems. At least partly because of the existence of these systems and the relative ease of accessing huge computerized medical databases, the use of databases in outcomes research has burgeoned [76]. These kinds of analyses are much faster and cheaper than conducting clinical trials. Databases can also be used to identify adverse events. Examples are comparisons of different antihypertensive agents and risk of stroke [77] and cyclooxygenase 2 (COX 2) inhibitors and risk of coronary heart disease [78]. In addition, databases likely represent a much broader population than the typical clinical trial, and can therefore complement clinical trial findings. This information can be useful as long as it is kept in mind that users and non-users of a medication are different and therefore have different characteristics.

Others [32, 79–81] have emphasized limitations of registry studies such as potential bias in treatment assignment, multiple comparisons, lack of standardization in collecting and reporting data, and missing data. Another weakness of prospective database registries is that they rely heavily on the validity of the model employed to analyze the data [82].

Lauer and D'Agostino note the high cost of clinical trials and argue that large databases may be able to substitute for trials that otherwise would not be conducted [83]. They also point out that existing registries and electronic health records can assist in conducting clinical trials. One such trial was the Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (TASTE), conducted in Scandinavia, which has extensive electronic health records [84].

There is no doubt that analyses of large databases can provide important information about disease occurrence and outcomes, as well as suggestions that certain therapies are preferable. As noted above, they can help to show that the results of clinical trials conducted in selected populations appear to apply in broader groups. Given their inherent chances for bias, however, they are no substitute for a randomized clinical trial in evaluating whether one intervention is truly better than another.

Cross-Over Designs

The cross-over design is a special case of a randomized control trial and has some appeal to medical researchers. The cross-over design allows each participant to serve as his own control. In the simplest case, namely the two period cross-over design, each participant will receive either intervention or control (A or B) in the first period and the alternative in the succeeding period. The order in which A and B are given to each participant is randomized. Thus, approximately half of the participants receive the intervention in the sequence AB and the other half in the sequence BA. This is so that any trend from first period to second period can be eliminated in the estimate of group differences in response. Cross-over designs need not be simple; they need not have only two groups. and there may be more than two periods [85, 86]. Depending on the duration of expected action of the intervention (for example, drug half-life), a wash-out period may be used between the periods.

The advantages and disadvantages of the two-period cross-over design have been described [19, 21, 86–89]. The appeal of the cross-over design to investigators is that it allows assessment of how each participant does on both A and B. Since each participant is used twice, variability is reduced because the measured effect of the intervention is the difference in an individual participant's response to intervention and control. This reduction in variability enables investigators to use smaller sample sizes to detect a specific difference in response. James et al. described 59 cross-over studies of analgesic agents. They concluded that if the studies had been designed using parallel or noncross-over designs, 2.4 times as many participants would have been needed [90]. Carriere showed that a three-period cross-over design is even more efficient than a two-period cross-over design [85].

In order to use the cross-over design, however, a fairly strict assumption must be made; the effects of the intervention during the first period must not carry over into the second period. This assumption should be independent of which intervention was assigned during the first period and of the participant response. In many clinical trials, such an assumption is clearly inappropriate, even if a wash-out is incorporated. If, for example, the intervention during the first period cures the disease, then the participant obviously cannot return to the initial state. In other clinical trials, the cross-over design appears more reasonable. If a drug's effect is to lower blood pressure or heart rate, then a drug-versus-placebo cross-over design might be considered if the drug has no carryover effect once the participant is taken off medication. Obviously, a fatal event and many disease complications cannot serve as the primary response variable in a cross-over trial.

Mills et al. [91] reviewed 116 reports of cross-over trials, which consisted of 127 individual trials. Reporting of key design and conduct characteristics was highly variable, making it difficult to discern whether optimal designs were followed.

As indicated in the International Conference on Harmonisation document E9, Statistical Principles for Clinical Trials [92], cross-over trials should be limited to those situations with few losses of study participants. A typical and acceptable cross-over trial, for example, might compare two formulations of the same drug in order to assess bioequivalence in healthy participants. Similarly, different doses may be used to assess pharmacologic properties. In studies involving participants who are ill or otherwise have conditions likely to change, however, cross-over trials have the limitations noted above.

Although the statistical method for checking the assumption of no period-treatment interaction was described by Grizzle [93], the test is not as powerful as one would like. What decreases the power of the test is that the mean response of the AB group is compared to the mean response of the BA group. However, participant variability is introduced in this comparison, which inflates the error term in the statistical test. Thus, the ability to test the assumption of no period-intervention interaction is not sensitive enough to detect important violations of the assumption unless many participants are used. The basic appeal of the cross-over design is to avoid between-participant variation in estimating the intervention effect, thereby requiring a smaller sample size. Yet the ability to justify the use of the design still depends on a test for carryover that includes between-participant variability. This weakens the main rationale for the cross-over design. Because of this insensitivity, the cross-over design is not as attractive as it at first appears. Fleiss et al. noted that even adjusting for baseline variables may not be adequate if inadequate time has been allowed for the participant to return to baseline status at the start of the second period [94]. Brown [19, 21] and Hills and Armitage [95] discourage the use of the cross-over design in general. Only if there is substantial evidence that the therapy has no carryover effects, and the scientific community is convinced by that evidence, should a cross-over design be considered.

Withdrawal Studies

A number of studies have been conducted in which the participants on a particular treatment for a chronic disease are taken off therapy or have the dosage reduced. The objective is to assess response to discontinuation or dose reduction. This design may be validly used to evaluate the duration of benefit of an intervention already known to be useful. For example, subsequent to the Hypertension Detection and Follow-up Program [96], which demonstrated the benefits of treating mild and moderate hypertension, several investigators withdrew a sample of participants with controlled blood pressure from antihypertensive therapy [97]. Participants were randomly assigned to continue medication, stop medication yet initiate nutritional changes, or stop medication without nutritional changes. After 4 years, only 5% of those taken off medication without nutritional changes remained normotensive and did not need the re-instatement of medication. This compared with 39% who were taken off medication yet instituted weight loss and reductions in salt

intake. Patients with severe chronic obstructive pulmonary disease (COPD) were prescribed a combination of tiotropium, salmeterol, and an inhaled glucocorticoid, fluticasone propionate for 6 weeks [98]. Because of the adverse effects of long term use of glucocorticoids, the investigators withdrew the fluticasone propionate over the subsequent 12 weeks. Despite a decrease in lung function, COPD exacerbations remained unchanged.

Withdrawal studies have also been used to assess the efficacy of an intervention that had not conclusively been shown to be beneficial in the long term. An early example is the Sixty Plus Reinfarction Study [99]. Participants doing well on oral anticoagulant therapy since their myocardial infarction, an average of 6 years earlier, were randomly assigned to continue on anticoagulants or assigned to placebo. Those who stayed on the intervention had lower mortality (not statistically significant) and a clear reduction in nonfatal reinfarction. A meta-analysis of prednisone and cyclosporine withdrawal trials (including some trials comparing withdrawal of the two drugs) in renal transplant patients has been conducted with graft failure or rejection as the response variables [100]. This meta-analysis found that withdrawal of prednisone was associated with increased risks of acute rejection and graft failure. Cyclosporine withdrawal led to an increase in acute rejection, but not graft failure. The Fracture Intervention Trial Long-term Extension (FLEX) assessed the benefits of continuing treatment with alendronate after 5 years of therapy [101]. The group that was randomized to discontinue alendronate had a modest increase in vertebral fractures but no increase in nonvertebral fractures.

One serious limitation of this type of study is that a highly selected sample is evaluated. Only those participants who physicians thought were benefiting from the intervention were likely to have been on it for several months or years. Anyone who had major adverse effects from the drug would have been taken off and, therefore, not been eligible for the withdrawal study. Thus, this design can overestimate benefit and underestimate toxicity. Another drawback is that both participants and disease states change over time.

If withdrawal studies are conducted, the same standards should be adhered to that are used with other designs. Randomization, blinding where feasible, unbiased assessment, and proper data analysis are as important here as in other settings.

Factorial Design

In the simple case, the factorial design attempts to evaluate two interventions compared to control in a single experiment [2–4, 102]. See Table 5.1.

Given the cost and effort in recruiting participants and conducting clinical trials, getting two (or more) experiments done at once is appealing. Examples of factorial designs are the Canadian transient ischemic attack study where aspirin and sulfinpyrazone were compared singly and together with placebo [103], the Third International Study of Infarct Survival (ISIS-3) that compared streptokinase, tissue plasminogen activator, and antistreplase plus aspirin plus heparin vs. aspirin

Table 5.1 Two-by-two factorial design

	Intervention X	Control	Marginals
Intervention Y	a	b	a + b
Control	c	d	c + d
Marginals	a + c	b + d	

Cell	*Intervention*
a	X + Y
b	Y + control
c	X + control
d	control + control

Effect of intervention X: a + c versus b + d

Effect of intervention Y: a + b versus c + d

alone [104], the Physicians' Health Study of aspirin and beta carotene [105], and the Women's Health Initiative (WHI) trial of hormone replacement, diet, and vitamin D plus calcium [106]. A review of analysis and reporting of factorial design trials [107] contains a list of 29 trials involving myocardial infarction and 15 other trials. Some factorial design studies are more complex than the 2 by 2 design, employing a third, or even a fourth level. It is also possible to leave some of the cells empty, that is, use an incomplete factorial design [108]. This was done in the Action to Control Cardiovascular Risk in Diabetes (ACCORD), which looked at intensive vs. less intensive glucose control plus either intensive blood pressure or lipid control [109]. This kind of design would be implemented if it is inappropriate, infeasible, or unethical to address every possible treatment combination. It is also possible to use a factorial design in a cross-over study [110].

The appeal of the factorial design might suggest that there really is a "free lunch." However, every design has strengths and weaknesses. A concern with the factorial design is the possibility of the existence of interaction between the interventions and its impact on the sample size. Interaction means that the effect of intervention X differs depending upon the presence or absence of intervention Y, or vice versa. It is more likely to occur when the two drugs are expected to have related mechanisms of action.

If one could safely assume there were no interactions, with a modest increase in sample size, two experiments can be conducted in one; one which is considerably smaller than the sum of two independent trials under the same design specifications. However, if one cannot reasonably rule out interaction, one should statistically test for its presence. As is true for the cross-over design, the power for testing for interaction is less than the power for testing for the main effects of interventions (cells a + c vs. b + d or cells a + b vs. c + d). Thus, to obtain satisfactory power to detect interaction, the total sample size must be increased. The extent of the increase depends on the degree of interaction, which may not be known until the end of the trial. The larger the interaction, the smaller the increase in sample size needed to detect it. If an interaction is detected, or perhaps only suggested, the comparison of intervention X would have to be done individually for intervention Y and its control (cell a vs. b and cell c vs. d). The power for these comparisons is obviously less than for the a + c vs. b + d comparison.

As noted, in studies where the various interventions either act on the same response variable or possibly through the same or similar mechanism of action, as with the presumed effect on platelets of both drugs in the Canadian transient ischemic attack study [103], interaction can be more of a concern. Furthermore, there may be a limited amount of reduction in the response variable that can be reasonably expected, restricting the joint effect of the interventions.

In trials such as the Physicians' Health Study [105], the two interventions, aspirin and beta carotene, were expected to act on two separate outcomes, cardiovascular disease and cancer. Thus, interaction was much less likely. But beta carotene is an antioxidant, and therefore might have affected both cancer and heart disease. It turned out to have no effect on either. Similarly, in the Women's Health Initiative [106], dietary and hormonal interventions may affect more than one disease process. There, diet had little effect on cancer and heart disease, but hormonal therapy had effects on heart disease, stroke, and cancer, among other conditions [111, 112].

In circumstances where there are two separate outcomes, e.g., heart disease and cancer, but one of the interventions may have an effect on both, data monitoring may become complicated. If, during the course of monitoring response variables it is determined that an intervention has a significant or important effect on one of the outcomes in a factorial design study, it may be difficult ethically, or even impossible, to continue the trial to assess fully the effect on the other outcome. Chapter 17 reviews data monitoring in more detail.

The factorial design has some distinct advantages. If the interaction of two interventions is important to determine, or if there is little chance of interaction, then such a design with appropriate sample size can be very informative and efficient. However, the added complexity, impact on recruitment and adherence, and potential adverse effects of "polypharmacy" must be considered. Brittain and Wittes [113] discuss a number of settings in which factorial designs might be useful or not, and raise several cautions. In addition to the issue of interaction, they note that less than full adherence to the intervention can exacerbate problems in a factorial design trial.

Group Allocation Designs

In group or cluster allocation designs, a group of individuals, a clinic or a community are randomized to a particular intervention or control [114–118]. The rationale is that the intervention is most appropriately or more feasibly administered to an entire group (for example, if the intervention consists of a broad media campaign). This design may also be better if there is concern about contamination. That is, when what one individual does might readily influence what other participants do. In the Child and Adolescent Trial for Cardiovascular Health, schools were randomized to different interventions [119]. Investigators randomized villages in a trial of vitamin A versus placebo on morbidity and mortality in children in India [120].

The Rapid Early Action for Coronary Treatment (REACT) trial involved ten matched pairs of cities. Within each pair, one city was randomly allocated to community education efforts aimed at reducing the time between symptoms of myocardial infarction and arrival at hospital [121]. Despite 18 months of community education, delay time was not different from that in the control cities. Communities have been compared in other trials [122, 123]. These designs have been used in cancer trials where a clinic or physician may have difficulty approaching people about the idea of randomization. The use of such designs in infectious disease control in areas with high prevalence of conditions such as tuberculosis and AIDS has become more common [124]. It should be noted that this example is both a group allocation design and a factorial design. Variations of group allocation, including cross-over and modification of cross-over, such as stepped wedge designs, where groups cross-over sequentially, rather than all at once, have been implemented [125, 126]. In the group allocation design, the basic sampling units and the units of analysis are groups, not individual participants. This means that the effective sample is substantially less than the total number of participants. Chapters 8 and 18 contain further discussions of the sample size determination and analysis of this design.

Hybrid Designs

Pocock [127] has argued that if a substantial amount of data is available from historical controls, then a hybrid, or combination design could be considered. Rather than a 50/50 allocation of participants, a smaller proportion could be randomized to control, permitting most to be assigned to the new intervention. A number of criteria must be met in order to combine the historical and randomized controls. These include the same entry criteria and evaluation factors, and participant recruitment by the same clinic or investigator. The data from the historical control participants must also be fairly recent. This approach, if feasible, requires fewer participants to be entered into a trial. Machin, however, cautions that if biases introduced from the non-randomized participants (historical controls) are substantial, more participants might have to be randomized to compensate than would be the case in a corresponding fully randomized trial [128].

Large, Simple and Pragmatic Clinical Trials

Advocates of large, simple trials maintain that for common medical conditions, it is important to uncover even modest benefits of intervention, particularly short-term interventions that are easily implemented in a large population. They also argue that an intervention is unlikely to have very different effects in different sorts of participants (i.e., subgroups). Therefore, careful characterization of people at

entry and of interim response variables, both of which add to the already considerable cost of trials, are unnecessary. The important criteria for a valid study are unbiased (i.e., randomized) allocation of participants to intervention or control and unbiased assessment of outcomes. Sufficiently large numbers of participants are more important than modest improvements in quality of data. The simplification of the study design and management allows for sufficiently large trials at reasonable cost. Examples of successfully completed large, simple trials are ISIS-3 [104], Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI) [129], Global Utilization of Streptokinase and Tissue Plasminogen Activator for Occluded Coronary Arteries (GUSTO) [130], a study of digitalis [131], the MICHELANGELO Organization to Assess Strategies in Acute Ischemic Syndromes (OASIS)-5 [132], and the Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia (TASTE) trial [84]. It should be noted that with the exception of the digitalis trial, these studies were relatively short-term. The questions addressed by these trials may be not only of the sort, "What treatment works better?" but "What is the best way of providing the treatment?" Can something shown to work in an academic setting be translated to a typical community medical care setting? Several have advocated conducting pragmatic or practical clinical trials. These kinds of trials, as noted in Chap. 3, are conducted in clinical practices, often far from academic centers. They address questions perceived as relevant to those practices [133–136]. Because of the broad involvement of many practitioners, the results of the trial may be more widely applied than the results of a trial done in just major medical settings. Thus, they may address a common criticism that the kinds of participants normally seen in academic centers, and therefore enrolled in many academic-based trials, are not the sort seen in typical clinical practices.

As indicated, these models depend upon a relatively easily administered intervention and an easily ascertained outcome. If the intervention is complex, requiring either special expertise or effort, particularly where adherence to protocol must be maintained over a long time, these kinds of studies are less likely to be successful. Similarly, if the response variable is a measure of morbidity that requires careful measurement by highly trained investigators, large simple or pragmatic trials are not feasible.

In recent years, the concept of comparative effectiveness research has become popular. Although trials comparing one agent against another have been conducted for many years, certain features of comparative effectiveness research should be mentioned. First, much of the research consists of other than clinical trials comparisons of interventions (e.g., use of databases as discussed in the sections above on nonrandomized control studies). In the clinical trial arena, much of the comparative effectiveness literature emphasizes studies done in collaboration with clinical practices (i.e., large, simple trials). They compare two or more interventions that are commonly used and involve outcome measures, including cost, that are of particular relevance to practitioners or to the participants [137].

It has also been pointed out that baseline characteristics may be useful for subgroup analysis. The issue of subgroup analysis is discussed more fully in Chap. 18. Although in general, it is likely that the effect of an intervention is

qualitatively the same across subgroups, exceptions may exist. In addition, important quantitative differences may occur. When there is reasonable expectation of such differences, appropriate baseline variables need to be measured. Variables such as age, gender, past history of a particular condition, or type of medication currently being taken can be assessed in a simple trial. On the other hand, if an invasive laboratory test or a measurement that requires special training is necessary at baseline, such characterization may make a simple or pragmatic trial infeasible.

The investigator also needs to consider that the results of the trial must be persuasive to others. If other researchers or clinicians seriously question the validity of the trial because of inadequate information about participants or inadequate documentation of quality control, then the study has not achieved its purpose.

There is no doubt that many clinical trials are too expensive and too cumbersome, especially multicenter ones. The advent of the large, simple trial or the pragmatic trial is an important step in enabling many meaningful medical questions to be addressed in an efficient manner. In other instances, however, the use of large numbers of participants may not compensate for reduced data collection and quality control. As always, the primary question being asked dictates the optimal design of the trial.

With increased understanding of genetic influences, the concept that interventions are likely to work similarly in all or at least most participants may no longer hold. There are differential effects of interventions in human epidermal growth factor receptor (HER-2) breast cancer, for example [138]. The concept of "personalized medicine" argues against the concept of large, simple trials and some have designed clinical trials to take advantage of biomarkers [139]. For most common conditions, however, we do not yet have the understanding required to implement personalized medicine, and large, simple trials will remain important for some time.

Studies of Equivalency and Noninferiority

Many clinical trials are designed to demonstrate that a new intervention is better than or superior to the control. However, not all trials have this goal. New interventions may have little or no superiority to existing therapies, but, as long as they are not materially worse, may be of interest because they are less toxic, less invasive, less costly, require fewer doses, improve quality of life, or have some other value to patients. In this setting, the goal of the trial would be to demonstrate that the new intervention is not worse, in terms of the primary response variable, than the standard by some predefined margin.

In studies of equivalency, the objective is to test whether a new intervention is equivalent to an established one. Noninferiority trials test whether the new intervention is no worse than, or at least as good as, some established intervention. Sample size issues for these kinds of trials are discussed in Chap. 8. It should also be noted that although the following discussion assumes one new intervention and one established intervention (the control), there is no reason why more complicated

Table 5.2 Noninferiority design assumptions

– Proper control arm
– Constancy over time and among participants
– Availability of data from prior studies of the control
– Assay sensitivity to demonstrate a true difference

designs involving multiple new interventions, for example, could not be implemented. This occurred in the Comparison of Age-Related Macular Degeneration Treatments Trials (CATT), where four groups (one standard therapy—monthly administration of intravitreal injections of ranibizumab—and three unproven therapies—as needed injections of ranibizumab and monthly and as needed injections of bevicizumab) were compared using a noninferiority design [140].

In equivalency and noninferiority trials, several design aspects need to be considered [141–148]. The control or standard treatment must have been shown conclusively to be effective; that is, truly better than placebo or no therapy. The circumstances under which the active control was found to be useful (i.e., similarity of populations, concomitant therapy, and dosage) ought to be reasonably close to those of the planned trial. These requirements also mean that the trials that demonstrated efficacy of the standard should be recent and properly designed, conducted, analyzed, and reported.

Table 5.2 shows the key assumptions for these trials. First, the active control that is selected must be one that is an established standard for the indication being studied and not a therapy that is inferior to other known ones. It must be used with the dose and formulation proven effective. Second, the studies that demonstrated benefit of the control against either placebo or no treatment must be sufficiently recent such that no important medical advances or other changes have occurred, and in populations similar to those planned for the new trial. Third, the evidence that demonstrated the benefits of the control must be available so that a control group event rate can be estimated. Fourth, the response variable used in the new trial must be sensitive to the postulated effects of the control and intervention. The proposed trial must be able to demonstrate "assay sensitivity," or the ability to show a difference if one truly exists. As emphasized in Chap. 8, the investigator must specify what she means by equivalence.

It cannot be shown statistically that two therapies are identical, as an infinite sample size would be required. Therefore, if the intervention falls sufficiently close to the standard, as defined by reasonable boundaries, the intervention is claimed to be "the same" as the control (in an equivalence trial) or no worse than the control (in a noninferiority trial). Selecting the margin of indifference or noninferiority, δ, is a challenge. Ideally, the relative risk of the new intervention compared to the control should be as close to 1 as possible. For practical reasons, the relative risk is often set in the range of 1.2–1.4. This means that in the worst case, the new intervention may be 20–40% inferior to standard treatment and yet be considered equivalent or noninferior. Some have even suggested that any new intervention could be approved by regulatory agencies as being noninferior to a standard control

intervention if it retains as least 50% of the control versus placebo effect. Further, there are options as to what 50% (or 40% or 20%) means. For example, one could choose either the point estimate from the control versus placebo comparison, or the lower confidence interval estimate of that comparison. Also, the choice of the metric or scale must be selected, such as a relative risk, or hazard ratio or perhaps an absolute difference. Of course, if an absolute difference that might seem reasonable with a high control group event rate is chosen, it might not seem so reasonable if the control group event rate turns out to be much lower than expected. This happened with a trial comparing warfarin against a new anticoagulant agent, where the observed control group event rate was less than that originally expected. Thus, with a predetermined absolute difference for noninferiority, the relative margin of noninferiority was larger than had been anticipated when the trial was designed [149].

It should be emphasized that new interventions are often hailed as successes if they are shown to be 20 or 25% better than placebo or a standard therapy. To turn around and claim that anything within a margin of 40 or 50% is equivalent to, or noninferior to a standard therapy would seem illogical. But the impact on sample size of seeking to demonstrate that a new intervention is at most 20% worse than a standard therapy, rather than 40%, is considerable. As is discussed in Chap. 8, it would not be just a twofold increase in sample size, but a fourfold increase if the other parameters remained the same. Therefore, all design considerations and implications must be carefully considered.

Perhaps even more than in superiority trials, the quality, the size and power of the new trial, and how well the trial is conducted, including how well participants adhere to the assigned therapy, are crucial. A small sample size or poor adherence with the protocol, leading to low statistical power, and therefore lack of significant difference, does not imply equivalence.

To illustrate the concepts around noninferiority designs, consider the series of trials represented in Fig. 5.5, which depicts estimates with 95% confidence intervals for the intervention effect.

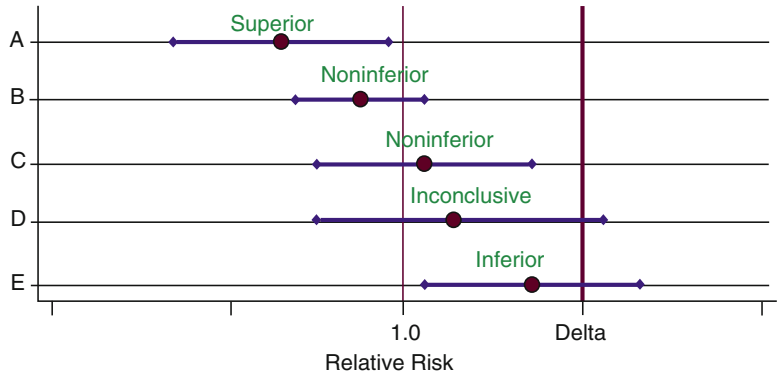


Fig. 5.5 Possible results of noninferiority trials

The heavy vertical line (labeled Delta) indicates the amount of worse effect of the intervention compared to the control that was chosen as tolerable. The thin vertical line indicates zero difference (a relative risk of 1). Trial A shows a new intervention that is superior to control (i.e. the upper confidence interval excludes zero difference). Trial B has an estimate of the intervention effect that is favorable but the upper limit of the confidence interval does not exclude zero. It is less than the margin of indifference, however, and thus meets the criterion of being noninferior. Trial C is also noninferior, but the point estimate of the effect is slightly in favor of the control. Trial D does not conclusively show superiority or noninferiority, probably because it is too small or there were other factors that led to low power. Trial E indicates inferiority for the new intervention.

As discussed above, the investigator must consider several issues when designing an equivalence or noninferiority trial. First, the constancy assumption that the control versus placebo effect has not changed over time is often not correct. This can be seen, for example, in two trials of the same design conducted back to back with essentially the same protocol and investigators, the PRAISE-1 and PRAISE-2 trials [57, 58] discussed in the section on Historical Controls and Databases. In PRAISE-1, the trial was stratified according to etiology, ischemic and non-ischemic heart failure. Most of the favorable effect of the drug on mortality was seen in the nonischemic stratum, contrary to expectation. To validate that subgroup result, PRAISE-2 was conducted in non-ischemic heart failure patients using the same design. In this second trial, no benefit of amlodipine was observed. The comparison of the placebo arms from PRAISE-1 and PRAISE-2 (Fig. 5.1), indicates that the two populations of nonischemic heart failure patients were at substantially different risk, despite being enrolled close in time, with the same entry criteria and same investigators. No covariate analysis could explain this difference in risk. Thus, the enrolled population itself is not constant, challenging the constancy assumption.

In addition, as background therapy changes, the effect of the control or placebo may also change. With more therapeutic options, the effect of one drug or intervention alone may no longer be as large as it was when placebo was the total background. Practice and referral patterns change.

Even if the data from prior trials of the selected control are available, the estimates of active control vs. placebo may not be completely accurate. As with all trials, effect of treatment depends at least partly on the sample of participants who were identified and volunteered for the study. The observed effect is not likely to reflect the effect exactly in some other population. It is also possible that the quality of the trials used to obtain the effect of the control may not have been very good. And of course, the play of chance may have affected the observed benefit.

Many of the assumptions about the active control group event rates that go into the design of a noninferiority or equivalence trial are unlikely to be valid. At the end of the trial, investigators obtain seemingly more precise estimates of the margin and imputed “efficacy,” when in fact they are based on a model that has considerable uncertainty and great care must be used in interpreting the results.

If I is the new intervention, C is the control or standard treatment, and P is placebo or not treatment, for the usual superiority trial, the goal is to show that the new intervention is better than placebo or no treatment, or that new intervention plus control is better than control alone.

$$I > P$$

$$I > C$$

$$I + C > C$$

For noninferiority trials, the margin of indifference, δ, is specified, where $I\text{-}C < \delta$. Efficacy imputation requires an estimate of the relative risk (RR) of the new intervention to control, RR(I/C) and of the control to placebo or no treatment, RR(C/P). Therefore, the estimated relative risk of the new intervention compared with placebo is

$$\mathrm{RR}(I/P) = \mathrm{RR}(I/C) \times \mathrm{RR}(C/P).$$

Rather than focus on the above assumption-filled model, an alternative approach might be considered. The first goal is to select the best control. This might be the one that, based on prior trials, was most effective. It might also be the one that the academic community considers as the standard of care, the one recommended in treatment guidelines, or the treatment that is most commonly used in practice. The selection will depend on the nature of the question being posed in the new trial. There might also be several possible best controls, all considered to be similar, as, for example, one of several beta blockers or statins. The choice might be influenced by regulatory agencies. The margin of noninferiority should use the data from the prior trials of the active control to get some estimate for initiating discussion but should not use it as a precise value. Once that estimate has been obtained, investigators, with input from others, including, as appropriate, those from regulatory agencies, should use their experience and clinical judgment to make a final determination as to what margin of noninferiority would support using a new intervention. These decisions depend on factors such as the severity of the condition being studied, the known risks of the standard or control intervention, the trade-offs that might be achieved with the new intervention, whether it is 50% or 20%, or some other relative risk, or an absolute difference, and the practicality of obtaining the estimated sample size. Having set the margin, effort must be on conducting the best trial, with as high participant adherence and complete follow-up as feasible. When the noninferiority trial has been completed, the attention should be given to the interpretation of trial results, keeping in mind the entirety of the research using the new intervention and the active control and the relevance of the findings to the specific clinical practice setting (see Chaps. 18 and 20).

Adaptive Designs

There is a great deal of interest in designs which are termed adaptive, but there are different designs that are adaptive and have different meanings of the term. Clinical trials have used forms of adaptive designs for many years. As discussed in Chap. 1, early phase studies have designs that allow for modifications as the data accrue. Many late phase trials are adaptive in the sense that the protocol allows for modification of the intervention in order to achieve a certain goal, typically using an interim variable. For example, trials of antihypertensive agents, with the primary response variable of stroke or heart disease, will allow, and even encourage, changes in dose of the agent, or addition or substitution of agent in order to reach a specified blood pressure reduction or level. A trial in people with depression changed antidepression drugs based on interim success or lack of success as judged by depression questionnaires [150]. Some have proposed re-randomizing either all participants or those failing to respond adequately to the first drug to other agents [151, 152].

Some trials, by design, will adjust the sample size to retain a desired power if the overall event rate is lower than expected, the variability is higher than planned, or adherence is worse than expected. In such cases, the sample size can be recalculated using the updated information (see Chap. 8). An event-driven adaptive design continues until the number of events thought necessary to reach statistical significance, given the hypothesized intervention effect, accumulates. In trials where time to event is the outcome of interest, the length of follow-up or the number of study participants, or both, may be increased in order to obtain the predetermined number of outcome events, In other adaptive designs, the randomization ratio may be modified to keep the overall balance between intervention and control arms level on some risk score (see Chap. 6).

Various designs are called response adaptive. Traditionally, if the effect of the intervention was less than expected, or other factors led to a less than desirable conditional power, the study either continued to the end without providing a clear answer or was stopped early for futility (see Chap. 17). Some studies, particularly where the outcome occurred relatively quickly, allowed for modification of the randomization ratio between intervention and control arm, depending on the response of the most recent participant or responses of all accumulated participants.

Because of concerns about inefficiencies in study design, several trend adaptive approaches have been developed. At the beginning of the trial, the investigator may have inadequate information about the rate at which the outcome variable will occur and be unable to make a realistic estimate of the effect of the intervention. Rather than continue to conduct an inappropriately powered trial or terminate early an otherwise well designed study, the investigator may wish to modify the sample size. After a trial is underway and better estimates become available, these trend adaptive approaches adjust sample size based on the observed trend in the primary outcome, in order to maintain the desired power. Trend adaptive designs require some adjustment of the analysis to assess properly the significance of the test statistic.

A criticism of these designs had been that they can introduce bias during the implementation of the adjustment. Some newer approaches, however, now allow for modifying sample size based on observed trends [153, 154]. They may also, however, provide sufficient information to allow people not privy to the accumulating data to make reasonable guesses as to the trend. See Chap. 18 for a further discussion of these methods.

Group sequential designs, in common use for many years, are also considered to be response adaptive in that they facilitate early termination of the trial when there is convincing evidence of benefit or harm. Response adaptive and trend adaptive designs will be considered further in Chaps. 17 and 18.

References

1. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.
2. Fisher RA. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.
3. Cochran WG, Cox GM *Experimental Designs* (2nd edition). New York: John Wiley and Sons, 1957.
4. Cox DR. *Planning of Experiments*. New York: John Wiley and Sons, 1958.
5. Bull JP. The historical development of clinical therapeutic trials. *J Chronic Dis* 1959;10:218–248.
6. Eliot MM. The control of rickets: preliminary discussion of the demonstration in New Haven. *JAMA* 1925;85:656–663.
7. Hill AB. Observation and experiment. *N Engl J Med* 1953;248:995–1001.
8. Macfarlane G. *Howard Florey: The Making of a Great Scientist*. Oxford: Oxford University Press, 1979, pp11–12.
9. Gocke DJ. Fulminant hepatitis treated with serum containing antibody to Australia antigen. *N Engl J Med* 1971;284:919.
10. Acute Hepatic Failure Study Group. Failure of specific immunotherapy in fulminant type B hepatitis. *Ann Intern Med* 1977;86:272–277.
11. Snow JB Jr, Kimmelman CP. Assessment of surgical procedures for Ménière's disease. *Laryngoscope* 1979;89:737–747.
12. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research* (4th edition). Malden, MA: Blackwell Publishing, 2002.
13. Brown BW, Hollander M. *Statistics: A Biomedical Introduction*. New York: John Wiley and Sons, 1977.
14. Feinstein AR. *Clinical Biostatistics*. St Louis: The C.V. Mosby Company, 1977.
15. MacMahon B, Trichopoulos D. *Epidemiology: Principles and Methods* (2nd edition). Lippincott Williams & Wilkins, 1996.
16. Lilienfeld DE, Stolley PD. *Foundations of Epidemiology* (3rd edition). New York: Oxford University Press, 1994.
17. Srivastava JN (ed.). *A Survey of Statistical Design and Linear Models*. Amsterdam: North-Hollard, 1975.
18. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. 1. Introduction and design. *Br J Cancer* 1976;34:585–612.
19. Brown BW Jr. Statistical controversies in the design of clinical trials—some personal views. *Control Clin Trials* 1980;1:13–27.
20. Pocock SJ. Allocation of patients to treatment in clinical trials. *Biometrics* 1979;35:183–197.
21. Brown BW Jr. The crossover experiment for clinical trials. *Biometrics* 1980;36:69–79.

22. Hennekens CH, Buring JC. *Epidemiology in Medicine*. SL Mayrent (ed.). Boston: Little, Brown, 1987.
23. Byar DP. Some statistical considerations for design of cancer prevention trials. *Prev Med* 1989;18:688–699.
24. Geller NL (ed.). *Advances in Clinical Trial Biostatistics*. New York: Marcel Dekker, 2003.
25. Piantadosi S. *Clinical Trials: A Methodologic Perspective* (2nd edition). New York: John Wiley and Sons, 2005.
26. Machin D, Day S, Green S. *Textbook of Clinical Trials* (2nd edition). West Sussex: John Wiley and Sons, 2006.
27. Green S, Benedetti J, Crowley J. *Clinical Trials in Oncology* (3rd edition). Boca Raton: CRC Press, 2012.
28. Hulley SB, Cummings SR, Browner WS, et al. *Designing Clinical Research* (4th edition). New York: Wolters Kluwer/Lippincott Williams & Wilkins, 2013.
29. Meinert CL. *Clinical Trials: Design, Conduct, and Analysis* (2nd edition). New York: Oxford University Press, 2012.
30. Cook TD, DeMets DL (eds). *Introduction to Statistical Methods for Clinical Trials*. Boca Raton: Chapman & Hall/CRC, Taylor & Francis Group, LLC, 2008.
31. Chow S-C, Shao J. *Statistics in Drug Research: Methodologies and Recent Developments*. New York: Marcel Dekker, 2002.
32. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. *Stat Med* 1984;3:361–373.
33. Gehan EA, Freireich EJ. Non-randomized controls in cancer clinical trials. *N Engl J Med* 1974;290:198–203.
34. Weinstein MC. Allocation of subjects in medical experiments. *N Engl J Med* 1974;291:1278–1285.
35. Byar DP, Simon RM, Friedewald WT, et al. Randomized clinical trials: perspectives on some recent ideas. *N Engl J Med* 1976;295:74–80.
36. Sapirstein W, Alpert S, Callahan TJ. The role of clinical trials in the Food and Drug Administration approval process for cardiovascular devices. *Circulation* 1994;89:1900–1902.
37. Hlatky MA. Perspective: Evidence-based use of cardiac procedures and devices. *N Engl J Med* 2004;350:2126–2128.
38. AMPLATZER® Septal Occluder. http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/DeviceApprovalsandClearances/Recently-ApprovedDevices/ucm083978.htm
39. St. Jude Amplatzer Atrial Septal Occluder (ASO): Safety communication—reports of tissue erosion. http://www.fda.gov/safety/medwatch/safetyinformation/safetyalertsforhumanmedicalproducts/ucm371202.htm
40. Chalmers TC, Matta RJ, Smith H, Kunzier AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1977;297:1091–1096.
41. Peto R. Clinical trial methodology. *Biomedicine* (Special issue) 1978;28:24–36.
42. Goldman L, Feinstein AR. Anticoagulants and myocardial infarction: the problems of pooling, drowning, and floating. *Ann Intern Med* 1979;90:92–94.
43. Grace ND, Muench H, Chalmers TC. The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology* 1966;50:684–691.
44. Sacks H, Chalmers TC, Smith H Jr. Randomized versus historical controls for clinical trials. *Am J Med* 1982;72:233–240.
45. Sacks HS, Chalmers TC, Smith H Jr. Sensitivity and specificity of clinical trials: randomized *v* historical controls. *Arch Intern Med* 1983;143:753–755.
46. Chalmers TC, Celano P, Sacks HS, Smith H Jr. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358–1361.
47. Ingelfinger FJ. The randomized clinical trial (editorial). *N Engl J Med* 1972;287:100–101.
48. Zelen M. A new design for randomized clinical trials. *N Engl J Med* 1979;300:1242–1245.

49. Anbar D. The relative efficiency of Zelen's prerandomization design for clinical trials. *Biometrics* 1983;39:711–718.
50. Ellenberg SS. Randomization designs in comparative clinical trials. *N Engl J Med* 1984;310:1404–1408.
51. Zelen M. Randomized consent designs for clinical trials: an update. *Stat Med* 1990;9:645–656.
52. Gehan EA. The evaluation of therapies: historical control studies. *Stat Med* 1984;3:315–324.
53. Lasagna L. Historical controls: the practitioner's clinical trials. *N Engl J Med* 1982;307:1339–1340.
54. Moertel CG. Improving the efficiency of clinical trials: a medical perspective. *Stat Med* 1984;3:455–465.
55. Pocock SJ. Letter to the editor. *Br Med J* 1977;1:1661.
56. Veterans Administration Cooperative Urological Research Group. Treatment and survival of patients with cancer of the prostate. *Surg Gynecol Obstet* 1967;124:1011–1017.
57. Packer M, O'Connor CM, Ghali JK, et al. for the Prospective Randomized Amlodipine Survival Evaluation Study Group. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. *N Engl J Med* 1996;335:1107–1114.
58. 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. *JACC:Heart Failure* 2013;1:308–314.
59. Havlik RJ, Feinleib M (eds.). *Proceedings of the Conference on the Decline in Coronary Heart Disease Mortality*. Washington, D.C.: NIH Publication No. 79-1610, 1979.
60. Health, United States, 2011, With Special Feature on Socioeconomic Status and Health. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. http://www.cdc.gov/nchs/data/hus/hus11.pdf, page 32, figure 3.
61. Health, United States, 2008, With Special Feature on the Health of Young Adults. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. http://www.cdc.gov/nchs/data/hus/hus08.pdf, page 37, figure 9.
62. Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 1975;231:360–381.
63. Castro KG, Ward JW, Slutsker L, et al. 1993 revised classification system for HIV infection and expanded surveillance case definitions for AIDS among adolescents and adults. *MMWR Recomm Rep,* December 18, 1992.
64. Current trends update: trends in AIDS diagnosis and reporting under the expanded surveillance definition for adolescents and adults—United States, 1993. *MMWR Weekly* 1994;43:826–831.
65. Rosenberg HM, Klebba AJ. Trends in cardiovascular mortality with a focus on ischemic heart disease: United States, 1950-1976. In Havlik R, Feinleib M (eds). *Proceedings of the Conference on the Decline in Coronary Heart Disease Mortality*. Washington, D.C.: NIH Publication No. 79-1610, 1979.
66. Morbidity and Mortality Chartbook on Cardiovascular, Lung, and Blood Diseases. National Heart, Lung, and Blood Institute, U.S. Department of Health and Human Services, Public Health Service. May 1994.
67. Centers for Disease Control and Prevention. International Classification of Diseases, (ICD-10-CN/PCS) Transition. http://www.cdc.gov/nchs/icd/icd10cm_pcs_impact.htm
68. Bailar JC III, Louis TA, Lavori PW, Polansky M. Studies without internal controls. *N Engl J Med* 1984;311:156–162.
69. Dustan HP, Schneckloth RE, Corcoran AC, Page IH. The effectiveness of long-term treatment of malignant hypertension. *Circulation* 1958;18:644–651.
70. Bjork S, Sannerstedt R, Angervall G, Hood B. Treatment and prognosis in malignant hypertension: clinical follow-up study of 93 patients on modern medical treatment. *Acta Med Scand* 1960;166:175–187.

71. Bjork S, Sannerstedt R, Falkheden T, Hood B. The effect of active drug treatment in severe hypertensive disease: an analysis of survival rates in 381 cases on combined treatment with various hypotensive agents. *Acta Med Scand* 1961;169:673–689.
72. Starmer CF, Lee KL, Harrell FE, Rosati RA. On the complexity of investigating chronic illness. *Biometrics* 1980;36:333–335.
73. Hlatky MA, Lee KL, Harrell FE Jr, et al. Tying clinical research to patient care by use of an observational database. *Stat Med* 1984;3:375–387.
74. Hlatky MA, Califf RM, Harrell FE Jr, et al. Clinical judgment and therapeutic decision making. *J Am Coll Cardiol* 1990;15:1–14.
75. Moon TE, Jones SE, Bonadonna G, et al. Using a database of protocol studies to evaluate therapy: a breast cancer example. *Stat Med* 1984;3:333–339.
76. Anderson C. Measuring what works in health care. *Science* 1994;263:1080–1082.
77. Klungel OH, Heckbert SR, Longstreth WT, et al. Antihypertensive drug therapies and the risk of ischemic stroke. *Arch Intern Med* 2001;161:37–43.
78. Graham DJ, Campen D, Hui R, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet* 2005;365:475–481.
79. Byar, DP. Why databases should not replace randomized clinical trials. *Biometrics* 1980;36:337–342.
80. Dambrosia JM, Ellenberg JH. Statistical considerations for a medical database. *Biometrics* 1980;36:323–332.
81. Sheldon TA. Please bypass the PORT. *Br Med J* 1994;309:142–143.
82. Mantel N. Cautions on the use of medical databases. *Stat Med* 1983;2:355–362.
83. Lauer MS, D'Agostino RB, Sr. The randomized registry trial—the next disruptive technology in clinical research? *N Engl J Med* 2013;369:1579–1581.
84. Fr□bert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-elevation myocardial infarction. *N Engl J Med* 2013:369:1587–1597; correction. *N Engl J Med* 2014;371:786.
85. Carriere KC. Crossover designs for clinical trials. *Stat Med* 1994;13:1063–1069.
86. Koch GG, Amara IA, Brown BW Jr, et al. A two-period crossover design for the comparison of two active treatments and placebo. *Stat Med* 1989;8:487–504.
87. Fleiss JL. A critique of recent research on the two treatment crossover design. *Control Clin Trials* 1989;10:237–243.
88. Woods JR, Williams JG, Tavel M. The two-period crossover design in medical research. *Ann Intern Med* 1989;110:560–566.
89. Louis TA, Lavori PW, Bailar JC III, Polansky M. Crossover and self-controlled designs in clinical research. *N Engl J Med* 1984;310:24–31.
90. James KE, Forrest WH, Jr, Rose RL. Crossover and noncrossover designs in four-point parallel line analgesic assays. *Clin Pharmacol Ther* 1985;37:242–252.
91. Mills EJ, Chan A-W, Wu P, et al. Design, analysis, and presentation of crossover trials. *Trials* 2009;10:27 doi:10.1186/1745-6215-10-27.
92. International Conference on Harmonisation: E9 Statistical principles for clinical trials. http://www.fda.gov/downloads/RegulatoryInformation/Guidances/UCM129505.pdf.
93. Grizzle JE. The two period change-over design and its use in clinical trials. *Biometrics* 1965;21:467–480.
94. Fleiss JL, Wallenstein S, Rosenfeld R. Adjusting for baseline measurements in the two-period crossover study: a cautionary note. *Control Clin Trials* 1985;6:192–197.
95. Hills M, Armitage P. The two-period cross-over clinical trial. *Br J Clin Pharmacol* 1979;8:7–20.
96. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the Hypertension Detection and Follow-Up Program. 1. Reduction in mortality of persons with high blood pressure, including mild hypertension. *JAMA* 1979;242:2562–2571.

97. Stamler R, Stamler J, Grimm R, et al. Nutritional therapy for high blood pressure-Final report of a four-year randomized controlled trial—The Hypertension Control Program. *JAMA* 1987;257:1484–1491.
98. Magnussen H, Disse B, Rodriguez-Roisin R, et al. Withdrawal of inhaled glucocorticoids and exacerbations of COPD. *N Engl J Med* 2014;371:1285–1294.
99. Report of the Sixty Plus Reinfarction Study Research Group. A double-blind trial to assess long-term oral anticoagulant therapy in elderly patients after myocardial infarction. *Lancet* 1980;316:989–994.
100. Kasiske BL, Chakkera HA, Louis TA, Ma JZ. A meta-analysis of immunosuppression withdrawal trials in renal transplantation. *J Am Soc Nephrol* 2000;11:1910–1917.
101. Black DM, Schwartz AV, Ensrud KE, et al, for the FLEX Research Group. Effects of continuing or stopping alendronate after 5 years of treatment: The Fracture Intervention Trial Long-term Extension (FLEX): a randomized trial. *JAMA* 2006;296:2927–2938.
102. Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomized controlled trials. *BMC Med Res Methodol* 2003;3:26doi:10.1186/1471-2288-3-26.
103. The Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. *N Engl J Med* 1978;299:53–59.
104. ISIS-3 (Third International Study of Infarct Survival) Collaborative Group. ISIS-3: a randomized study of streptokinase vs plasminogen activator vs anistrephase and of aspirin plus heparin vs aspirin alone among 41,299 cases of suspected acute myocardial infarction. *Lancet* 1992;339:753–770.
105. Stampfer MJ, Buring JE, Willett W, et al. The 2 x 2 factorial design: its application to a randomized trial of aspirin and carotene in U.S. physicians. *Stat Med* 1985;4:111–116.
106. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials* 1998;19:61–109.
107. McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: a systematic review. *JAMA* 2003;289:2545–2553.
108. Byar DP, Herzberg AM, Tan W-Y. Incomplete factorial designs for randomized clinical trials. *Stat Med* 1993;12:1629–1641.
109. Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in Type 2 diabetes. *N Engl J Med* 2008;358:2545–2559.
110. Fletcher DJ, Lewis SM, Matthews JNS. Factorial designs for crossover clinical trials. *Stat Med* 1990;9:1121–1129.
111. 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.
112. The Women's Health Initiative Steering Committee. Effects of conjugated equine estrogen in postmenopausal women with hysterectomy. *JAMA* 2004;291:1701–1712.
113. Brittain E, Wittes J. Factorial designs in clinical trials: the effects of non-compliance and subadditivity. *Stat Med* 1989;8:161–171.
114. Hayes RJ, Moulton LH. *Cluster Randomized Trials: A Practical Approach*. Chapman & Hall/CRC, Taylor & Francis Group, 2009.
115. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–914.
116. Armitage P. The role of randomization in clinical trials. *Stat Med* 1982;1:345–352.
117. Simon R. Composite randomization designs for clinical trials. *Biometrics* 1981; 37:723–731.
118. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol* 1978;108:100–102.
119. Zucker DM, Lakatos E, Webber LS, et al. Statistical design of the Child and Adolescent Trial for Cardiovascular Health (CATCH): implications of cluster randomization. *Control Clin Trials* 1995;16:96–118.
120. Vijayaraghavan K, Radhaiah G, Prakasam BS, et al. Effect of massive dose vitamin A on morbidity and mortality in Indian children. *Lancet* 1990;336:1342–1345.

121. Luepker RV, Raczynski JM, Osganian S, et al. Effect of a community intervention on patient delay and emergency medical service use in acute coronary heart disease: the Rapid Early Action for Coronary Treatment (REACT) trial. *JAMA* 2000;284:60–67.
122. Farquhar JW, Fortmann SP, Flora JA, et al. Effects of community-wide education on cardiovascular disease risk factors. The Stanford Five-City Project. *JAMA* 1990;264:359–365.
123. Gail MH, Byar DP, Pechacek TF, Corle DK, for COMMIT Study Group. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation. *Control Clin Trials* 1992;13:6–21.
124. Sismanidis C, Moulton LH, Ayles H, et al. Restricted randomization of ZAMSTAR: a 2x2 factorial cluster randomized trial. *Clin Trials* 2008;5:316–327.
125. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;28:182–191.
126. Woertman W, de Hoop E, Moerbeek M, et al. Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J Clin Epidemiol* 2013;66:752–758.
127. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis* 1976;29:175–188.
128. Machin D. On the possibility of incorporating patients from nonrandomising centres into a randomised clinical trial. *J Chronic Dis* 1979;32:347–353.
129. Gruppo Italiano per lo Studio della Streptochinasi nell' Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. *Lancet* 1986; *i*:397–402.
130. The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med* 1993;329:673–682; correction *N Engl J Med* 1994;331:277.
131. The Digitalis Investigation Group. Rationale, design, implementation, and baseline characteristics of patients in the DIG Trial: a large, simple, long-term trial to evaluate the effect of digitalis on mortality in heart failure. *Control Clin Trials* 1996;17:77–97.
132. MICHELANGELO OASIS 5 Steering Committee. Design and rationale of the MICHELANGELO Organization to Assess Strategies in Acute Ischemic Syndromes (OASIS)-5 trial program evaluating fondaparinux, a synthetic factor Xa inhibitor, in patients with non-ST-segment elevation acute coronary syndromes. *Am Heart J* 2005;150:1107–1114.
133. 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.
134. March JS, Silva SG, Compton S, et al. The case for practical clinical trials in psychiatry. *Am J Psychiatry* 2005;162:836–846.
135. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol* 2009;62:464–475.
136. Johnson KE, Tachibana C, Coronado GD, et al. Research Methods & Reporting: A guide to research partnerships for pragmatic trials. *BMJ* 2014;349:g6826 doi:10.1136/bmj.g6826.
137. Mailankody S, Prasad V. Perspective: Comparative effectiveness questions in oncology. *N Engl J Med* 2014; 370:1478–1481.
138. Ross JS, Slodkowska EA, Symmans WF, et al. The HER-2 receptor and breast cancer: ten years of targeted anti-HER-2 therapy and personalized medicine. *Oncologist* 2009;14:320–368.
139. Lai TL, Lavori PW, Shih MC, Sikic BI. Clinical trial designs for testing biomarker-based personalized therapies. *Clin Trials* 2012;9:141–154.
140. The CATT Research Group. Ranibizumab and bevacizumab for neovascular age-related macular degeneration. *N Engl J Med* 2011;364:1897–1908.
141. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982;3:345–353.
142. Hung JHM, Wang SJ, Tsong Y, et al. Some fundamental issues with non-inferiority testing in active controlled trials. *Stat Med* 2003;30:213–225.

143. Fleming TR. Current issues in non-inferiority trials. *Stat Med* 2008;27:317–332.
144. D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Stat Med* 2003;22:169–186.
145. 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.
146. Mulla SM, Scott IA, Jackevicius CA, You JJ, Guyatt GH. How to use a noninferiority trial. *JAMA* 2012;308:2605–2611.
147. Schumi J, Wittes JT. Through the looking glass: understanding non-inferiority. *Trials* 2011;12:106 doi:10.1186/1745-6215-12-106.
148. DeMets DL, Friedman L. Some thoughts on challenges for noninferiority study designs. *Therapeutic Innovation & Regulatory Science* 2012;46:420–427.
149. SPORTIF Executive Steering Committee for the SPORTIF V Investigators. Ximelagatran vs warfarin for stroke prevention in patients with nonvalvular atrial fibrillation: a randomized trial. *JAMA* 2005;293:690–698.
150. Trivedi MH, Rush AJ, Wisniewski SR, et al, for the STAR*D Study Team. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry* 2006;163:28–40.
151. Murphy SA, Oslin DW, Rush AJ, Zhu J, for MCATS. Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders (Perspective). *Neuropsychopharmacology* 2007;32:257–262.
152. Lavori PW, Dawson D. Improving the efficiency of estimation in randomized trials of adaptive treatment strategies. *Clin Trials* 2007;4:297–308.
153. Levin GP, Emerson SC, Emerson SS. Adaptive clinical trial designs with prespecified rules for modifying the sample size: understanding efficient types of adaptation. *Stat Med* 2013;32:1259–1275.
154. Mehta C. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: a different perspective. *Stat Med* 2013;32:1276–1279.