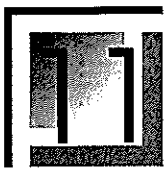


Designing an Experiment: Clinical Trials II

Deborah Grady,
Steven R. Cummings,
and Stephen B. Hulley

In the last chapter, we discussed the randomized, blinded trial: how to select participants, measure baseline variables, randomize, and apply the intervention. In this chapter, we describe how to maximize follow-up and adherence to the protocol, measure the outcome, and analyze the results. Clinical trials are very different from observational studies in that something is done to participants, and this chapter addresses the need for monitoring the results during the trial. The chapter ends by reviewing some alternatives to the classic randomized trial.

■ FOLLOW-UP AND ADHERENCE TO THE PROTOCOL

If a substantial number of study participants do not receive the study intervention, do not adhere to the protocol, or are lost to follow-up, the results of the trial are likely to be underpowered or biased. Strategies for **maximizing follow-up and adherence** are outlined in Table 11.1.

The effect of the intervention (and the power of the trial) is reduced to the degree that participants do not receive it. The investigator should try to choose a study drug or behavioral intervention that is easy to apply or take and is well tolerated. Adherence is likely to be poor if a behavioral intervention requires hours of practice by participants. Drugs that can be taken in a single daily dose are the easiest to remember and therefore preferable. The protocol should include provisions that will enhance adherence, such as instructing participants to take the pill at a standard point in the morning routine and giving them pill containers labeled with the day of the week.

There is also a need to consider how best to **measure adherence** to the intervention, using such approaches as self-report, pill counts, automated pill dispensers, and serum or urinary metabolite levels. This information can identify participants who are not complying, so that the investigator can help explain the finding if there is no difference between groups at the end.

Adherence to study visits and measurements can be enhanced by discussing what is involved in the study before consent is obtained, by scheduling the visits at a time that is convenient and with enough staff to prevent waiting, by calling the participant the day before each visit, and by reimbursing travel expenses and other out-of-pocket costs.

Failure to follow trial participants and measure the outcome of interest can result in biased results, diminished credibility of the findings, and decreased

■ TABLE 11.1
Maximizing Follow-up and Adherence to the Protocol

Principle	Example
Choose subjects who are likely to be adherent to the intervention and protocol	Require completion of two or more comprehensive visits before randomization
	Exclude those who are nonadherent in a prerandomization run-in period
	Exclude those who are likely to move or be noncompliant
Make the intervention easy	Use a single tablet rather than two
Make study visits convenient and enjoyable	Schedule visits often enough to maintain close contact but not frequently enough to be tiresome
	Schedule visits at night or on weekends, or collect information by e-mail
	Have adequate staff to prevent waiting
	Provide reimbursement for travel
	Establish personal relationships with subjects
Make study measurements painless and interesting	Choose noninvasive, informative tests that are not otherwise available
	Provide test results of interest to participants and appropriate counseling
Encourage subjects to continue in the trial	Never discontinue subjects from follow-up for protocol violations, adverse events, or side effects
	Send participants birthday and holiday cards
	Send newsletters and e-mail messages
	Emphasize the scientific importance of adherence and follow-up
Find subjects who are lost to follow-up	Pursue contacts of subjects, and use a tracking service.

statistical power. For example, a trial of nasal calcitonin spray to reduce the risk of osteoporotic fractures reported that treatment reduced fracture risk by 36% (1). However, about 60% of those randomized were lost to follow-up, and it was not known if fractures had occurred in these participants. Because the overall number of fractures was small, even a few fractures in the participants lost to follow-up could have altered the findings of the trial. This uncertainty diminished the credibility of the study findings (2).

Even if participants violate the protocol or discontinue the trial intervention, they should be followed so that their outcomes can be used in intention-to-treat analyses. In many trials, participants who violate the protocol by enrolling in another trial, discontinue the study intervention, or report adverse effects are

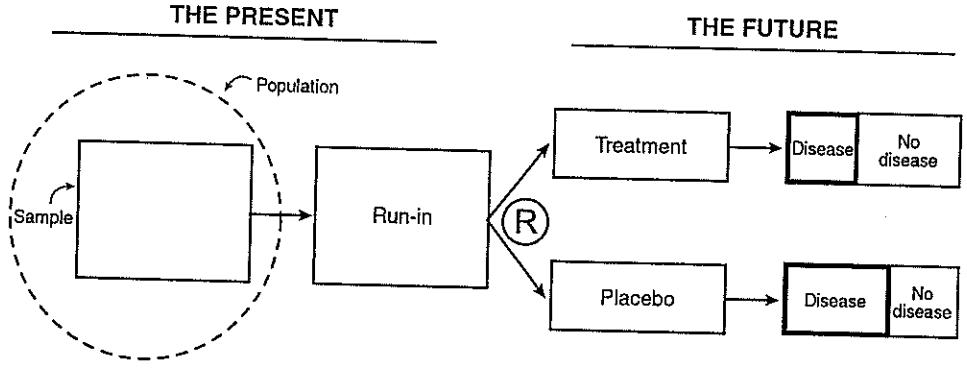
discontinued from follow-up; this can result in biased or uninterpretable results. Consider, for example, a drug that causes a symptomatic side effect that frequently results in discontinuation of the study medication. If participants who discontinue study medication are not followed for the outcome, the rate of events in the active treatment group will be biased downward. This bias can have a serious effect on the main findings if the side effect is associated with the main outcome.

Some strategies for achieving complete **follow-up** are similar to those discussed for cohort studies (Chapter 7). At the outset of the study, participants should be informed of the importance of follow-up and investigators should record the name, address, and telephone number of one or two close acquaintances who will always know where the participant is. In addition to enhancing the investigator's ability to assess vital status, this may give him access to proxy outcome measures that can be obtained by telephone from those who absolutely refuse to come for a visit at the end. In the HERS trial, 89% of the women returned for the final clinic visit, another 8% had a final telephone contact for outcome ascertainment, and information on vital status was determined for every single participant by using phone contact, registered letters, contacts with close relatives, and a tracking service (3).

The design of the trial should make it as easy as possible for participants to adhere to the intervention and complete all follow-up visits and measurements. Long and stressful visits can deter some participants from attending. Participants are more likely to return for visits that involve noninvasive tests, such as electron beam computed tomography of the heart, than for invasive tests such as coronary angiography. Collecting follow-up information by phone or electronic means may improve adherence for participants who find visits difficult. On the other hand, participants may lose interest in a trial if there are not some social or interpersonal rewards for participation. Participants may tire of study visits that are scheduled monthly, and they may lose interest if visits only occur annually. Follow-up is also improved by making the trial experience positive and enjoyable for study participants: designing trial measurements and procedures to be painless and interesting; performing tests that would not otherwise be available; providing results of tests to participants (if the result will not influence outcomes); sending newsletters, e-mail notes of appreciation, holiday, and birthday cards; giving inexpensive gifts; and developing strong personal relationships with study staff.

Two design aspects that are specific to trials may improve adherence and follow-up: screening visits before randomization and a run-in period. Asking participants to attend one or two **screening visits** before randomization may exclude participants who find that they cannot complete such visits. The trick here is to set the hurdles for entry into the trial high enough to exclude those who will later be nonadherent, but not high enough to exclude participants who will turn out to have satisfactory adherence.

A **run-in period** may be a useful design for increasing the proportion of study participants who adhere to the intervention and follow-up procedures. During the baseline period, all participants are placed on placebo. A specified time later (usually a few weeks), those who have complied with the intervention are randomized blindly to continue taking the placebo or to begin taking the active drug. Excluding nonadherent participants before randomization in this fashion may increase the power of the study and permit a better estimate of the full effects of intervention. It is not clear, however, that a placebo run-in is more effective than the requirement that participants complete one or more screening visits before randomization.



■ **FIGURE 11.1**

In a randomized trial preceded by a run-in period to test compliance, the investigator (a) selects a sample from the population, (b) measures baseline variables, (c) randomizes the participants, (d) applies interventions, (d) follows up the cohorts, (e) measures outcome variables.

A variant of the placebo run-in design shown in Fig. 11.1 is the use of the active drug rather than the placebo for the run-in period. In addition to increasing adherence among those who enroll, an active drug run-in is designed to select participants who tolerate and respond to the intervention. The response of an intermediary variable (i.e., a surrogate that lies between the intervention and the outcome) is used as the criterion for randomization. In a trial of an antiarrhythmic drug's effect on mortality, for example, the investigator might randomize only those participants whose arrhythmias are satisfactorily suppressed without undue side effects (4). This design maximizes power by increasing the proportion of the intervention group that is responsive to the intervention. It also improves generalizability by mimicking the clinician's tendency to continue using a drug only when he sees evidence that it is working. When those who do not tolerate or do not respond to an intervention are excluded from a trial, the results may not be generalizable to those excluded.

It is also possible that the rate of adverse effects among those enrolled will underestimate the rate among all who are placed on the intervention. A trial of the effect of carvedilol on mortality in patients with congestive heart failure used a 2-week active run-in period. During the run-in, 17 people had worsening congestive heart failure and seven died (5). These people were not randomized in the trial, and these adverse effects of drug treatment were not included as outcomes.

■ MEASURING THE OUTCOME

In choosing the outcome measure the investigator often must balance **clinical relevance** with **feasibility** and **cost**.

Clinical versus Surrogate Outcomes

Clinically relevant measures, such as death, myocardial infarction, hospital admission, and quality of life, are the most meaningful outcomes of trials. Surrogate markers for risk of the outcome, such as cholesterol for risk of coronary heart disease (CHD), are used when the testing of a new treatment is at a relatively early stage and when resources are too limited to permit a large study with clinical outcomes (Chapter 10). At a minimum, surrogate markers must be biologically plausible and associated with the outcome of interest; for example, bone density

is commonly used as a surrogate marker for risk of fracture because the low bone density of osteoporosis has been shown to be associated with an increased risk of fracture. This does not prove, however, that treatments that cause favorable changes in the surrogate marker will produce better clinical outcomes. Distressingly, there are many instances where trials using surrogate markers for clinical outcomes have produced misleading results. For example, several studies showed that ventricular arrhythmias increase risk for death among patients with myocardial infarction. Subsequent trials also showed that certain drugs could suppress ventricular arrhythmia (the surrogate outcome). Unfortunately, the Cardiac Arrhythmia Suppression Trial (CAST) demonstrated that even though these drugs reduced the frequency of serious arrhythmia, the mortality rate was higher among treated patients (4).

Statistical Characteristics

The outcome measure should be one that can be assessed accurately and precisely. An example of an outcome that meets these criteria is a newborn baby's weight; an example of one that does not is the presence of a congenital learning disability, a behavioral variable that represents the ill-defined end of a continuum.

Continuous outcome variables have the advantage over dichotomous ones of enhancing the power of the study, thus permitting a smaller sample size. In Chapter 6 a study with birth weight as a continuous outcome variable requires less than half the sample size needed for a study in which the outcome is the proportion of newborns who weigh less than 2,500 grams. Unfortunately, birth weight as a continuous variable is much less clinically relevant because differences in birth weight among those babies who weigh more than 2,500 grams—about 90% of all babies—may not be related to any clinical problem.

If a dichotomous outcome is unavoidable, power depends more on the number of events than on the overall number of participants (6). In the HERS trial, for example, power was not determined by the 2,763 women in the trial, but by the 348 who experienced the primary outcome—nonfatal myocardial infarction or CHD death (2). A dichotomous outcome that was more common, such as all acute coronary syndromes (nonfatal myocardial infarction, CHD death, and hospitalization for unstable angina), which occurred in 568 women, could be tested with proportionally greater power.

Number of Outcome Variables

It is often desirable to have several outcome variables that measure different aspects of the phenomena of interest. In the HERS trial, CHD events were chosen as the primary end point. Nonfatal myocardial infarction, CHD death, revascularization, hospitalization for unstable angina and congestive heart failure, stroke and transient ischemic attack, venous thromboembolic events, and all-cause mortality were all assessed and adjudicated to provide a more detailed description of the cardiovascular effects of hormone therapy (3). However, a **single primary end point** was designated for the purpose of planning the sample size and duration of the study and to avoid the problems of interpreting tests of multiple hypotheses (Chapter 5).

Adjudication of Outcomes

Most self-reported outcomes, such as history of stroke or a participant report of quitting smoking, are not 100% accurate. Self-reported outcomes that are important to the trial should be confirmed if possible. Occurrence of disease, such as a

stroke, is generally adjudicated by (a) creating clear criteria for the outcome (new neurologic deficit with corresponding lesion on computed tomography or magnetic resonance imaging scan), (b) collecting the clinical documents needed to make the assessment (discharge summaries and radiology reports), and (c) having experts review each potential case and judge whether the criteria for the diagnosis have been met. Those who collect the information and adjudicate the cases must be blinded to the treatment assignment.

Adverse Effects

The investigator should include outcome measures that will detect the occurrence of **adverse effects** that may result from the intervention. Revealing whether the beneficial effects of an intervention outweigh the adverse ones is a major goal of most clinical trials, even those that test apparently innocuous treatments like a health education program. Adverse effects may range from relatively minor symptoms such as rash or flulike episodes, to serious and fatal complications. The investigator should consider the problem that both the nature of the end point and the sample size requirements for detecting adverse effects may be different from those for detecting benefits. Unfortunately, rare side effects will usually be impossible to detect no matter how large the trial and are discovered (if at all) only after an intervention is in widespread clinical use.

In the early stages of testing a new treatment when potential adverse effects are unclear, investigators should ask broad, open-ended questions about all types of potential adverse effects. In large trials, assessment and coding of all potential adverse events can be very expensive and time-consuming, with a low yield of important results. Investigators should consider strategies for minimizing this burden while preserving an adequate assessment of potential harms of the intervention. For example, common events, such as respiratory infections and gastrointestinal upset, may be assessed in a subset of the participants or for a limited time. Important potential adverse events or effects that are expected because of previous research may be more accurately and efficiently assessed by specific queries. For example, since rhabdomyolysis is a reported side effect of treatment with statins, the signs and symptoms of myositis should be queried in any trial of a new statin. When data from a trial will be used to apply for approval of a new drug, the trial design must satisfy regulatory expectations for reporting adverse events. (See "Good Clinical Practices" on the U.S. FDA website.)

■ ANALYZING THE RESULTS

Statistical analysis of the primary hypothesis of a clinical trial is generally straightforward. If the outcome is dichotomous, the simplest approach is to compare the proportions in the study groups using a chi-squared test. When the outcome is continuous a *t* test may be used, or a nonparametric alternative if the outcome is not normally distributed. In most clinical trials the duration of follow-up is different for each participant, necessitating the use of survival time methods. More sophisticated statistical models such as Cox proportional hazards analysis can accomplish this and at the same time adjust for chance maldistributions of baseline confounding variables. The technical details of when and how to use these methods are described elsewhere (7).

Two important issues that should be considered in the analysis of clinical trial results are the primacy of the intention-to-treat analytic approach and the ancillary role for subgroup analyses.

Intention-to-Treat Analysis

For analysis, the investigator must decide what to do with "**cross-overs**," participants assigned to the active treatment group who do not get treatment or discontinue it and those assigned to the control group who end up getting active treatment. An analysis done by **intention-to-treat** compares outcomes between the study groups with every participant analyzed according to his randomized group assignment, regardless of whether he received the assigned intervention. Intention-to-treat analyses may underestimate the full effect of the treatment, but they guard against more important causes of biased results in clinical trials.

An alternative to intention-to-treat is to analyze only those who comply with the intervention. It is common, for example, to perform "**per protocol**" analyses that include only those participants in both groups who took more than 80% of their assigned study medication or only those who are "evaluable" (i.e., took a certain proportion of study medication, completed a certain proportion of visits, and had no other protocol violations). This seems reasonable because participants can only be affected by an intervention they actually receive. The problem arises, however, that participants who adhere to the study treatment may be different from those who drop out in ways that are related to the outcome. In the Postmenopausal Estrogen-Progestin Interventions Trial (PEPI), 875 postmenopausal women were randomly assigned to four different estrogen or estrogen plus progestin regimens and placebo (8). Among women assigned to the unopposed estrogen arm, 30% had discontinued treatment after 3 years because of endometrial hyperplasia, which is a precursor of endometrial cancer. If these women are eliminated in a per protocol analysis, an association of estrogen therapy and endometrial cancer may be missed.

The major disadvantage of the intention-to-treat approach is that participants who choose not to take the assigned intervention will nevertheless be included in the estimate of the effects of that intervention. Thus substantial discontinuation or cross-over between treatments will cause intention-to-treat analyses to underestimate the magnitude of the effect of treatment. For this reason, results of trials are often evaluated with both intention-to-treat and per protocol analyses. If both analyses produce similar results, this increases confidence in the conclusions of the trial. If they differ, results of the intention-to-treat analyses generally predominate because they preserve the value of randomization and, unlike per protocol analyses, can only bias the estimated effect in the conservative direction (favoring H_o). The results can only be analyzed in both ways if follow-up measures are completed regardless of whether participants adhere to treatment.

Subgroup analyses are defined as comparisons between randomized groups in a subset of the trial cohort. These analyses have a mixed reputation because they are easy to misuse and can lead to wrong conclusions. With proper care, however, they can provide useful ancillary information and expand the inferences that can be drawn from a clinical trial. To preserve the value of randomization, subgroups should be defined by measurements that were made before treatment was started. For example, a trial of alendronate to prevent osteoporotic fractures found that the drug decreased risk of fracture by 14% among women with low bone density. Preplanned analyses by subgroups of bone density measured at baseline revealed that the treatment was effective (36% reduction in fracture risk; $P < 0.01$) among women whose bone density was more than 2.5 standard deviations below normal. In contrast, treatment was ineffective in women with higher bone density at baseline (9). It is important to note that the value of randomization

is preserved in each of the subgroups: the fracture rate among women randomized to alendronate is compared with the rate among women randomized to placebo in each subgroup—those with very low bone density (defined by measurements made before randomization) and those with higher bone density.

Subgroup analyses are prone, however, to producing misleading results for several reasons. Subgroups are, by definition, smaller than the entire trial population, and there may not be sufficient power to find important differences; investigators should avoid claiming that a drug "was ineffective" in a subgroup when the problem might reflect insufficient power to find an effect. Investigators often examine results in a large number of subgroups, increasing the likelihood of finding a different effect of the intervention in one subgroup by chance. Optimally, planned subgroup analyses should be defined before the trial begins and the number of subgroups analyzed should be reported with the results of the study. A conservative approach is to require that claims about different responses in subgroups be supported by statistical evidence that there is an interaction between treatment and the subgroup characteristic. For example, the study of alendronate found a significant interaction ($P = 0.01$) between baseline bone density and the effect of treatment on risk of fractures, supporting the conclusion that alendronate works in women with osteoporosis but not in women with higher bone density.

Subgroup analyses based on postrandomization factors do not preserve the value of randomization and often produce misleading results. Per protocol analyses limited to subjects who adhere to the randomized treatment are examples of this type of subgroup analysis.

■ MONITORING CLINICAL TRIALS

Why monitor a clinical trial? An important difference between clinical trials and observational studies is that in a clinical trial something is being done to the participants. For ethical reasons, investigators must assure that participants not be exposed to a harmful intervention, denied a beneficial intervention, or continued in a trial if the research question cannot possibly be answered.

The most pressing reason to monitor clinical trials is to make sure that the intervention does not turn out unexpectedly to be harmful. If the harm outweighs any benefits, the trial should be stopped. Second, if an intervention is more effective than was estimated when the trial was designed, then benefit can be observed early in the trial. When clear benefit has been proved, it may be unethical to continue the trial and delay offering the intervention to participants on placebo and to others who could benefit. Third, if there is no possibility of answering the research question, it may be unethical to continue participants in a trial that requires time and effort and that may cause some discomfort or risk. If a clinical trial is scheduled to continue for 5 years but after 4 years there is little difference in the rate of outcome events in the treated and untreated groups, then the "conditional power" (the likelihood of answering the research question given the results, thus far) becomes very small and consideration should be given to stopping the trial. In addition, because clinical trials are expensive, stopping the trial as soon as the question is answered saves money.

The research question might be answered by other trials before a given trial is finished. It is desirable to have more than one trial that provides evidence concerning a given research question, but if definitive evidence becomes available during a trial, the investigator should consider stopping.

How interim monitoring will occur should be considered in the planning of any clinical trial. Guidelines and procedures for monitoring should be detailed

■ TABLE 11.2
Monitoring a Clinical Trial

- Elements to monitor
  - Recruitment
  - Adherence
  - Randomization
  - Blinding
  - Follow-up
  - Important variables
    - Outcomes
    - Adverse effects
    - Potential confounders
- Who will monitor
  - Trial investigators if small trial with minor hazards
  - Independent data monitoring board otherwise
- Changes in the protocol that can result from monitoring
  - Terminate the trial
  - Modify the trial
    - Stop one arm of the trial
    - Add new measurements necessary for safety monitoring
    - Discontinue high-risk participants
  - Extend the trial in time
  - Enlarge the trial sample
- How often to monitor
  - Often enough to meet the goals of monitoring
  - Only when there is substantial new data
- Statistical methods for monitoring

in writing before the study begins. Items to include in these guidelines are outlined in Table 11.2.

Stopping a trial should always be a carefully weighted decision that balances ethical responsibility to the participants and the advancement of scientific knowledge. Whenever a trial is stopped early, the chance to provide more conclusive results will be lost. The decision is often complex, and potential risks to participants must be weighed against possible benefits. Thus it is important that the committee that monitors the trial include physicians, participant advocates, biostatisticians, and persons experienced in conducting trials. These experts are normally outsiders who are not involved in the trial, and therefore have no personal or financial interest in its continuation.

Statistical tests of significance provide important but not conclusive information for stopping a trial. Trends over time should be evaluated for consistency, effects on related outcomes should be evaluated for consistency, and the impact of stopping the study early on the credibility of the findings should be carefully considered (Example 11.1).

There are many statistical methods for monitoring the interim results of a trial. Analyzing the results of a trial repeatedly is a form of multiple testing and thus increases the probability of a type I error. For example, if $\alpha = 0.05$ is used for each test and the results of a trial are analyzed four times during the trial and again at the end, the probability of making a type I error is increased to about 14% (14). To address this problem, statistical methods for interim monitoring generally decrease the α for each test (α_i) so that the overall $\alpha = 0.05$. There are multiple approaches to deciding how to "spend α" (Appendix 11.1).

Example 11.1 Trials That Have Been Stopped Early

Canadian Atrial Fibrillation Anticoagulation Study, CAFA (10): Atrial fibrillation is a risk factor for stroke and embolic events. The CAFA study was a double-blind, randomized, placebo-controlled trial to evaluate the efficacy of warfarin in decreasing the rate of stroke, systemic embolism, or intracerebral or fatal bleeding in patients with nonrheumatic atrial fibrillation. The study was designed to enroll 660 patients and follow them on therapy for 3.5 years. During the trial (after 383 patients had been randomized and followed for a mean of 1.2 years), the results of two other randomized trials were reported showing a significant decrease in embolic events and a low rate of major bleeding events in patients with atrial fibrillation treated with warfarin. The Steering Committee of the CAFA decided that the evidence of benefit with warfarin was sufficiently compelling to stop the trial without preliminary examination of the data.

Cardiac Arrhythmia Suppression Trial, CAST (4): The occurrence of ventricular premature depolarizations in survivors of myocardial infarction is a risk factor for sudden death. The CAST evaluated the effect of antiarrhythmic therapy (encainide, flecainide, or moricizine) in patients with asymptomatic or mildly symptomatic ventricular arrhythmia after myocardial infarction on risk for sudden death. During an average of 10 months of follow-up, the participants treated with active drug had a higher total mortality (7.7% versus 3.0%) and a higher rate of death from arrhythmia (4.5% versus 1.5%) than those assigned to placebo. The trial was planned to continue for 5 years but was stopped after 18 months.

Coronary Drug Project, CDP (11,12): The CDP was a randomized, blinded trial to determine if five different cholesterol-lowering interventions (conjugated estrogen 5.0 mg/day; estrogen 2.5 mg/day; clofibrate 1.8 g/day; dextrothyroxine 6.0 mg/day; niacin 3.0 g/day) reduced the 5-year mortality rate. The CDP enrolled 8,341 men with myocardial infarction who were followed for at least 5 years. With an average of 18 months of follow-up, the high-dose estrogen arm was stopped due to an excess of nonfatal myocardial infarction (6.2% compared with 3.2%) and venous thromboembolic events (3.5% compared with 1.5%). This decision was reinforced by the fact that high-dose estrogen was also associated with testicular atrophy, gynecomastia, breast tenderness, and decreased libido. At the same time, dextrothyroxine was stopped in the subgroup of men who had demonstrated frequent premature ventricular beats on their baseline electrocardiogram because the death rate in this subgroup was 38.5% compared with 11.5% in the same subgroup receiving placebo. Dextrothyroxine therapy was stopped in all subjects shortly thereafter due to an excess mortality rate in the treated group. Two years before the planned end of the study, the 2.5-mg-dose estrogen arm was also stopped because there was no evidence of any beneficial effect and an increased risk of venous thromboembolic events among treated men.

Physicians Health Study (13): The Physicians Health Study was a randomized trial of the effect of aspirin (325 mg every other day) on cardiovascular mortality. The trial was stopped after 4.8 years of the planned 8-year follow-up. There was a significant reduction in myocardial infarction in the treated group (relative risk for nonfatal MI = 0.56), but the number of cardiovascular disease deaths in each group was equal. The rate of cardiovascular disease deaths observed in the study was much lower than expected (88 after 4.8 years of follow-up versus the 733 expected), and the trial was stopped because the conditional power to detect a favorable impact of aspirin therapy on cardiovascular mortality had fallen to a very low level).

■ ALTERNATIVES TO THE RANDOMIZED BLINDED TRIAL

Other Randomized Designs

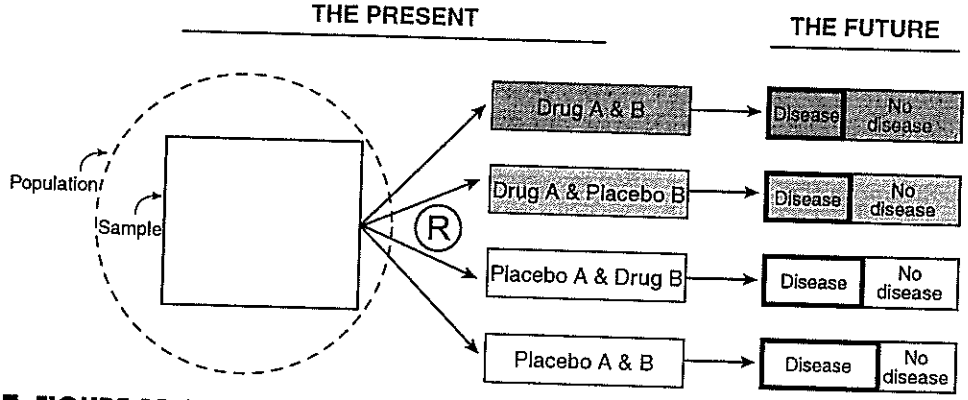
There are a number of variations on the classic randomized trial that may be useful when the circumstances are right.

The **factorial design** aims to answer two separate research questions in a single cohort of participants (Fig. 11.2). A good example is the Physicians' Health Study which was designed to test the effect of aspirin on myocardial infarction and of beta-carotene on cancer (15). The participants were randomly assigned to four groups, but each of the two hypotheses was tested by comparing two halves of the study cohort. First, all those on aspirin are compared with all those on aspirin placebo (disregarding the fact that half of each of these groups received beta-carotene); then all those on beta-carotene are compared with all those on beta-carotene placebo (now disregarding the fact that half of each of these groups received aspirin). The investigator has two complete trials for the price of one.

The factorial design is very efficient. The chief limitation is the possibility of interactions between the treatments and outcomes. In the example noted earlier, any influence of beta-carotene on myocardial infarction would alter the outcome for half of the participants receiving aspirin, reducing power and confusing interpretation. Factorial designs can actually be used to study such interactions, but these trials are more complicated and difficult to implement, large sample sizes are required, and the results can be hard to interpret. In clinical research, the best role for factorial designs is in studying two relatively unrelated research questions.

Randomization of matched pairs is a strategy for balancing baseline confounding variables that requires selecting pairs of subjects who are matched on important factors like age and sex, then randomly assigning one member of each pair to each study group. A particularly attractive version of this design can be used when the circumstances permit a contrast of treatment and control effects in two parts of the same individual at the same time. In the Diabetic Retinopathy Study, for example, each participant had one eye randomly assigned to photocoagulation treatment while the other served as a control (16).

Group or cluster randomization requires that the investigator randomly assign



■ **FIGURE 11.2**

In a factorial randomized trial, the investigator (a) selects a sample from the population; (b) measures baseline variables; (c) randomly assigns two active interventions and their controls to four groups, as shown; (d) applies interventions; (e) follows up the cohorts; (f) measures outcome variables.

naturally occurring groups or clusters of participants to the study groups rather than assign individuals. A good example is a trial that enrolled players on 120 college baseball teams, randomly allocated half of the teams to an intervention to encourage cessation of spit-tobacco use, and observed a significantly lower rate of spit-tobacco use among players on the teams that received the intervention (17). Applying the intervention to groups of people may be more feasible and cost-effective than treating individuals one at a time, and it may better address research questions about the effects of public health programs in the population. Some interventions, such as a low-fat diet, are difficult to implement in one member of a family but not in others. Similarly, participants who receive a transferable intervention may discuss this advice with acquaintances who have been assigned to the control group. For example, a clinician in a group practice who is randomly assigned to an educational intervention is very likely to discuss this intervention with his colleagues. A disadvantage of cluster randomization is the fact that sample size estimation and analysis are more complicated (18).

Nonrandomized Between-Group Designs

Trials that compare groups that have not been randomized are far less satisfactory than randomized trials in controlling for the influence of confounding variables. Analytic methods can adjust for baseline factors that are unequal in the two study groups, but this strategy does not deal with the problem of unmeasured confounding. Chalmers has reviewed the findings of randomized and nonrandomized studies of the same research question (19); the apparent benefits of intervention were much greater in the nonrandomized studies, even after adjusting statistically for differences in baseline variables. This and other analyses (20) indicate that the problem of confounding in nonrandomized clinical studies can be serious and that it may not be fully removed by statistical adjustment.

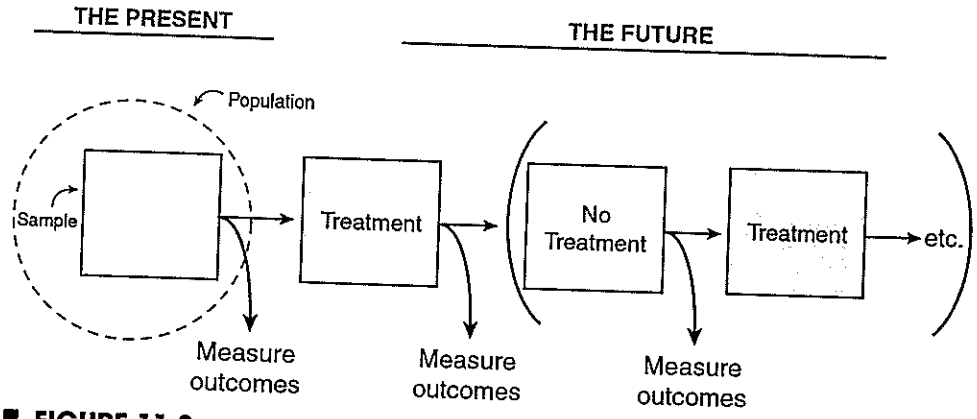
Sometimes subjects are allocated to the study groups by a pseudo-random mechanism. For example, every other subject (or every subject with an even hospital record number) may be assigned to the treatment group. Such designs sometimes offer logistic advantages, but the predictability of the study group assignment permits the investigator to tamper with it by manipulating the sequence or eligibility of new subjects.

Sometimes subjects are assigned to study groups by the investigator according to certain clinical criteria. For example, diabetic patients may be allocated to receive either insulin four times a day or long-acting insulin once a day according to their willingness to accept four daily injections. The problem is that those willing to take four injections per day might be more compliant with other health advice, and this might be the cause of any observed difference in the outcomes of the two treatment programs.

Nonrandomized designs are sometimes chosen in the mistaken belief that they are more ethical. In fact, studies are only ethical if they are designed well enough to have a reasonable likelihood of producing the correct answer to the research question, and randomized designs are more likely to lead to a conclusive result than nonrandomized designs. Moreover, the ethical basis for any trial is the uncertainty as to whether the intervention will be beneficial or harmful, an uncertainty termed **equipoise** that must exist if the trial needs to be done at all.

Within-Group Designs

Designs that do not include randomization can be useful options for some types of questions (Fig. 11.3). In a **time-series design,** each participant serves as his own control to evaluate the effect of treatment. This means that innate characteris-



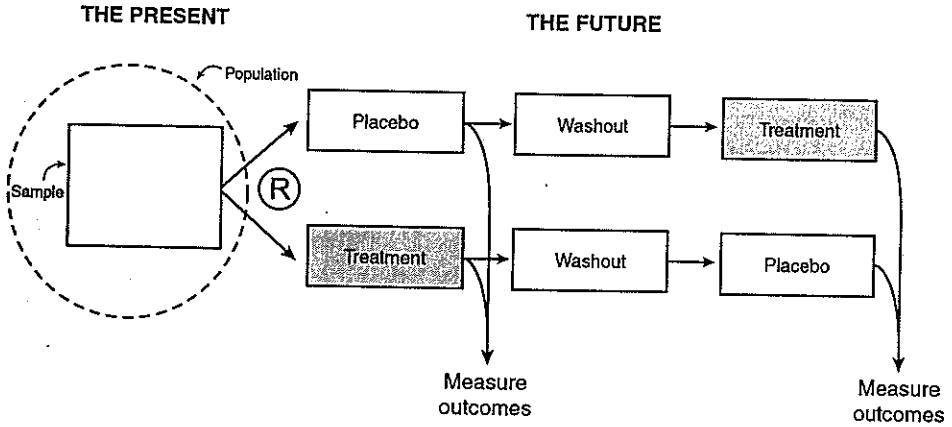
■ **FIGURE 11.3**

In a time series study, the investigator (a) selects a sample from the population, (b) measures baseline and outcome variables, (c) applies intervention to the whole cohort, (d) follows up the cohort, (e) measures outcome variables again, (f) (optional) removes intervention and measures outcome variables again.

tics such as age, sex, and genetic factors are not merely balanced (as they are in between-group studies) but actually eliminated as confounding variables.

The major disadvantage of within-group designs is the **lack of a concurrent control group.** The apparent efficacy of the intervention might be due to **learning effects** (participants do better on follow-up cognitive function tests because they learned from the baseline test), **regression to the mean** (participants who were selected for the trial because they had high blood pressure at baseline are found to have lower blood pressure at follow-up simply due to random variation in blood pressure), or **secular trends** (upper respiratory infections are less frequent at follow-up because the trial started during flu season). Within-group designs sometimes use a strategy of repeatedly starting and stopping the treatment. If repeated onset and offset of the intervention produces similar patterns in the outcome, this provides strong support that these changes are due to the treatment. This approach is only useful when the outcome variable responds rapidly and reversibly to the intervention (the effect of alcohol intake on HDL-cholesterol level, for example).

The **cross-over design** has features of both within- and between-group designs (Fig. 11.4). Half of the participants are randomly assigned to start with the control period and then switch to active treatment; the other half do the opposite. This approach (or the Latin square for more than two treatment groups) permits between-group as well as within-group analyses. The advantages of this design are substantial: It further minimizes the potential for confounding because each participant serves as his own control and substantially increases the statistical power of the trial so that it needs fewer participants. However, the disadvantages are also substantial: a doubling of the duration of the study, and the added complexity of analysis and interpretation created by the problem of **carryover effects.** A carryover effect is the residual influence of the intervention on the outcome during the period after it has been stopped. Blood pressure may not return to baseline levels for months after a course of diuretic treatment, for example. To reduce the carryover effect, the investigator can introduce an untreated **"washout"** period with the hope that the outcome variable will return to normal before starting the next intervention, but it is difficult to know whether all carryover effects have been eliminated. In general, crossover studies are only a good



■ **FIGURE 11.4**

In the cross-over randomized trial, the investigator (a) selects a sample from the population, (b) measures baseline variables, (c) randomizes the participants, (d) applies interventions, (e) measures outcome variables, (f) allows washout period to reduce carryover effect, (g) applies intervention to former placebo group, (h) measures outcome variables again.

choice when the number of study subjects is limited and carryover effects are judged not to be a problem.

■ TRIALS FOR FDA APPROVAL OF NEW THERAPIES

Many trials are done to test the effectiveness and safety of new treatments that might be considered for approval by the U.S. Food and Drug Administration (FDA) or another national regulatory body for marketing. Trials are also done to determine whether drugs that have FDA approval for one condition might be approved for the treatment or prevention of other conditions.

The FDA publishes detailed and updated guidelines for how such trials should be conducted. (Search for "FDA" on the Web.) Guidelines also cover European and international regulations for approval (called International Committee on Harmonization [ICH] guidelines, which can be found on the Web). It is wise for investigators who conduct these trials to seek specific training in "**Good Clinical Practices,**" which are guidelines available on the FDA Web site for the conduct of clinical trials by investigators and staff who enroll and treat participants.

Trials of new treatments are generally described by stage. This system refers to an orderly progression in the testing of a new treatment, from experiments in animals (**preclinical**) and initial unblinded and uncontrolled administration to a few human volunteers to test the safety of the treatment (**phase I**), to relatively small randomized blinded trials that test the effect of a range of doses on side effects and surrogate measurements of the clinical outcome that is the target of the treatment (**phase II**), to randomized trials large enough to test the hypothesis that the treatment improves the targeted condition (such as blood pressure) or reduces the risk of disease (such as stroke) with acceptable safety (**phase III**) (Table 11.3). **Phase IV** refers to large studies (which may or may not be randomized trials) conducted after a drug is approved. These studies are often conducted (and financed) by marketing departments of pharmaceutical companies with the goals

■ **TABLE 11.3**
Stages in Testing New Therapies

Preclinical	Studies in Cell Culture and Animals
Phase I	Unblinded, uncontrolled studies in a few volunteers to test safety
Phase II	Relatively small randomized, controlled, blinded trials to test tolerability and different intensity or dose of the intervention on surrogate outcomes
Phase III	Relatively large randomized, controlled, blinded trials to test the effect of the therapy on clinical outcomes
Phase IV	Large trials or observational studies conducted after the therapy has been approved by the FDA to assess the rate of serious side effects and evaluate additional therapeutic uses

of assessing the rate of serious side effects when used in very large populations and identifying additional uses of the drug that might be approved by the FDA.

■ DECIDING TO DO A TRIAL

In general, research questions should be answered with randomized trials if feasible. The major advantage of a randomized trial is its potential for controlling the influence of confounding variables, thus providing more conclusive answers. For some research questions, a trial may be faster and less expensive than observational studies, particularly when the outcome variable is continuous and responds rapidly to the intervention. For example, it is difficult to demonstrate the relationship between dietary fat and serum cholesterol in an observational study (because of errors in measuring in the dietary variable) but relatively easy to do so in a trial. For some research questions a trial is clearly necessary to control for confounding and to make sure that the benefit outweighs the risk. For example, observational studies have consistently found that people who take beta-carotene have a lower risk of cancer, but four large clinical trials have failed to find a benefit (21); the findings of the observational studies may be due to confounding because people who take vitamins may be more health conscious than those who do not.

However, trials are usually time-consuming and expensive, and often expose participants to discomfort or risk. Therefore they should not be performed until enough is known about the intervention to suggest that a definitive trial is possible. Such information includes definition of the *exact* intervention (therapy, counseling, surgical procedure, or drug dose, duration, and route), the likely benefit of the intervention (to allow estimation of sample size and duration of the trial), and the likely adverse effects of the intervention (to allow adequate safety protection for participants). A clinical trial should not be undertaken when, because of the absence of randomization, blinding, or sufficient numbers of participants, it is unlikely to provide a conclusive answer.

SUMMARY

1. If a substantial number of study participants **do not receive** the study intervention, **do not adhere** to the protocol, or are **lost to follow-up,** the results of the trial are likely to be underpowered, biased, or uninterpretable.
2. **Clinically relevant measures,** such as death, myocardial infarction, hospital admission, and quality of life, are the most meaningful outcomes of trials. To the extent possible, the investigator should include outcome measures that will detect the occurrence of **adverse effects** that may result from the intervention.
3. **Intention-to-treat** analyses are the primary approach to take advantage of the control over confounding provided by randomization. **Per protocol** analyses, a secondary approach that provides an estimate of the effect size in adherent subjects, should be interpreted with caution.
4. With proper care, **subgroup analyses** can provide useful ancillary information and expand the inferences that can be drawn from a clinical trial. To preserve the value of randomization, subgroups should be defined by measurements that were made before treatment was started, and analyses should compare outcomes between subsets of randomly assigned study groups.
5. An important difference between clinical trials and observational studies is that in a clinical trial, *something is being done to the participants.* **Interim monitoring** during a trial should make sure that participants are not exposed to a harmful intervention, denied a beneficial intervention, or continued in a trial if the research question cannot possibly be answered.
6. There are several variations on the randomized trial design that can substantially increase efficiency under the right circumstances:
   a. The **factorial design** allows two independent trials to be carried out for the price of one.
   b. **Matched-pair** randomization balances baseline confounding variables.
   c. **Group randomization** permits efficient studies of naturally occurring clusters.
   d. **Time-series designs** have a single (non-randomized) group with outcomes compared within each subject during periods of different interventions.
   e. **Cross-over designs** may control for confounding and minimize the required sample size if carryover effects are not a problem.

EXERCISES

1. a. Continuing with the research question, "Does mild vitamin D deficiency cause hip fractures in the elderly," briefly outline an **experiment** designed to answer the question. Contrast the advantages and disadvantages of the experimental approach compared with observational designs.

b. List some strategies for making your experiment more cost-effective.

c. There is evidence that if vitamin D has an effect on hip fractures, it may do so by improving muscle strength. What ideas does this give you for designing a more effective study?

References

1. Chestnut CH, Silverman S, Andriano K, et al. A randomized trial of nasal spray salmon calcitonin in postmenopausal women with established osteoporosis: the PROOF study. *Am J Med,* in press.

■ AP
Ou

2. Cummings S, Chapurlat R. What PROOF proves about calcitonin and clinical trials. *Am J Med*, in press.
3. Hulley S, Grady D, Bush T, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *JAMA* 1998;280:605–13.
4. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. *N Engl J Med* 1989;321:406–12.
5. Pfeffer M, Stevenson L. Beta-adrenergic blockers and survival in heart failure. *N Engl J Med* 1996;334:1396–7.
6. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409–20.
7. Friedman LM, DeMets DL, Furberg C. *Fundamentals of clinical trials*, 3rd ed. St. Louis: Mosby-Year Book, 1996.
8. Writing Group for the PEPI Trial. Effects of estrogen or estrogen/progestin regimens on heart disease risk factors in postmenopausal women. *JAMA* 1995;273:199–208.
9. Cummings S, Black D, Thompson D, et al. Effect of alendronate on risk of fracture in women with low bone density but without vertebral fractures: results from the Fracture Intervention Trial. *JAMA* 1998;280:2077–82.
10. Laupacis A, Connolly SJ, Gent M, et al. How should results from completed studies influence ongoing clinical trials? The CAFA Study experience. *Ann Intern Med* 1991;115:818–22.
11. The Coronary Drug Project. Initial findings leading to modifications of its research protocol. *JAMA* 1970;214:1303–13.
12. The Coronary Drug Project. Findings leading to discontinuation of the 2.5-mg day estrogen group. The coronary Drug Project Research Group. *JAMA* 1973;226:652–7.
13. Findings from the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 1988;318:262–4.
14. Armitage P, McPherson C, Rowe B. Repeated significance tests on accumulating data. *J R Stat Soc* 1969;132A:235–44.
15. Hennekens C, Eberlein K. A randomized trial of aspirin and beta-carotene among U.S. physicians. *Prev Med* 1985;14:165–8.
16. Diabetic Retinopathy Study Research Group. Preliminary report on effects of photocoagulation therapy. *Am J Ophthalmol* 1976;81:383–96.
17. Walsh M, Hilton J, Masouredis C, et al. Smokeless tobacco cessation intervention for college athletes: results after 1 year. *Am J Pub Health* 1999;89:228–34.
18. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–14.
19. Chalmers T, Celano P, Sacks H, et al. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358–61.
20. Pocock S. Current issues in the design and interpretation of clinical trials. *Br Med J* 1985;296:39–42.
21. Marshall J. Beta-carotene: a miss for epidemiology. *J Natl Cancer Inst* 1999;91:2068–9.

■ APPENDIX 11.1 Interim Monitoring of Trial Outcomes

Interim monitoring of trial results is a form of multiple testing, and thus increases the probability of a type I error. To address this problem, α for each test (α_i) is generally decreased so that the overall $\alpha = 0.05$. There are multiple statistical methods for decreasing α.

One of the easiest to understand is the Bonferroni method, where $\alpha_i = \alpha/N$ if N is the total number of tests performed. For example, if the overall α is 0.05 and five tests will be performed, α_i for each test is 0.01. This method has several

disadvantages, however. It requires using an equal threshold for stopping the trial at any interim analysis. Most investigators would rather use a lower threshold for stopping a trial earlier rather than later and the Bonferroni approach results in a very low α for the final analysis. In addition, this approach is too conservative because it assumes that each test is independent. For these reasons, Bonferroni is not generally used.

A commonly used method suggested by O'Brien and Fleming (1) uses a very small initial α_i, then gradually increases it such that α_i for the final test is close to the overall α. O'Brien-Fleming provide methods for calculating α_i if the investigator chooses the number of tests to be done and the overall α. At each test, $Z_i = Z^* (N_i)^{1/2}$, where $Z_i = Z$ value for the ith test; Z^* is determined so as to achieve the overall significance level; N is the total number of tests planned and i is the ith test. For example, for five tests and overall $\alpha = 0.05$, $Z^* = 2.04$; the initial $\alpha = 0.00001$ and the final $\alpha_5 = 0.046$. This method is unlikely to lead to stopping a trial very early unless there is a striking difference in outcome between randomized groups (as was the case in CAST [4]). In addition, this method avoids the awkward situation of getting to the end of a trial and accepting the null hypothesis even though the P value is substantially less than 0.05.

A major drawback to the preceding methods is that the number of tests and the proportion of data to be tested must be decided before the trial starts. In some trials, additional interim tests are necessary when important trends occur. Lan and DeMets (2) developed a method using a specified α-spending function that provides continuous stopping boundaries. The α_i at a particular time (or after a certain proportion of outcomes) is determined by the function and by the number of previous "looks." Using this method, neither the number of "looks" nor the proportion of data to be analyzed at each "look" must be specified before the trial. Of course, for each additional interim analysis conducted, the final α is lower.

A different set of statistical methods based on curtailed sampling techniques suggests termination of a trial if future data are unlikely to change the conclusion. The multiple testing problem is irrelevant because the decision is based only on estimation of what the data will show at the end of the trial. A common approach is to compute the conditional probability of rejecting the null hypothesis at the end of the trial, based on the accumulated data. First, conditional power is calculated assuming that H_o is true (i.e., that any future outcomes in the treated and control groups will be equally distributed). Second, H_a is assumed to be true (i.e., that outcomes will be distributed unequally in the treatment and control groups). The effect size is usually assumed to be the same as that used to calculate the sample size but it can be made somewhat more extreme. If the conditional power to reject the null hypothesis under either of these two assumptions is low, the null hypothesis is not likely to be rejected and the trial might be stopped.

References

1. O'Brien P, Fleming T. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549–56.
2. DeMets D, Lan G. The alpha spending function approach to interim data analyses. *Cancer Treat Res* 1995;75:1–27.