10 Designing an Experiment: Clinical Trials I

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

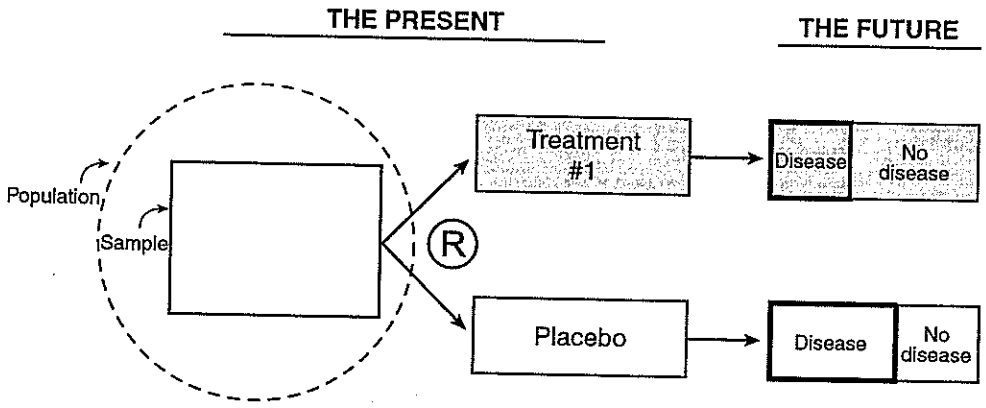
In clinical trials, the investigator applies a treatment (termed **intervention**) and observes the effect on an outcome. The major advantage of a trial over an observational study is the ability to demonstrate causality. In particular, **randomly assigning** the intervention can eliminate the influence of confounding variables, and **blinding** its administration can eliminate the possibility that the observed effects of the intervention are due to other treatments or to biased ascertainment.

However, clinical trials are generally expensive, are time-consuming, address a narrow clinical question, and sometimes expose participants to potential harm. For these reasons, trials are best reserved for relatively mature research questions, when observational studies and other lines of evidence suggest that an intervention might be effective but stronger evidence is required as the basis for practice guidelines. Not every research question is amenable to the clinical trial design. For example, it is not feasible to study whether drug treatment of high-LDL cholesterol in children will prevent heart attacks many decades later. But the principles of evidence-based medicine require clinical trial evidence whenever possible.

This chapter focuses on the classic **randomized blinded trial** (Fig. 10.1), addressing the selection of participants, measurement of baseline variables, randomization, and choice of intervention and control. In the next chapter we will come to outcomes, analysis, and other trial designs.

SELECTING THE PARTICIPANTS

Selection of the participants begins with deciding what kind of participants to study and how to go about recruiting them. Chapter 3 discussed how to specify entry criteria defining a target population that is appropriate to the research question and an accessible population that is practical to study, how to design an efficient and scientific approach to selecting participants, and how to recruit them. Here are some further points that are especially relevant to clinical trials.



■ **FIGURE 10.1**

In a randomized trial, the investigator (a) selects a sample from the population, (b) measures baseline variables, (c) randomizes the participants, (d) applies interventions (one should be a blinded placebo, if possible), (e) follows up the cohort, (f) measures outcome variables (blindly, if possible) and analyzes the results.

Define Entry Criteria

In a clinical trial, inclusion and exclusion criteria have the joint goal of identifying an important population for whom a statistically significant impact of the intervention on the outcome is feasible and likely. This means that the entry criteria should optimize the rate of the primary outcome, the expected effectiveness of the active treatment, the generalizability of findings from the trial, the ease of recruitment, and the likelihood of compliance with treatment and follow-up.

If the outcome of interest is a rare event, such as breast cancer, it is usually necessary to recruit participants who have a high risk of the outcome to reduce the sample size and follow-up time to feasible levels. On the other hand, narrowing the inclusion criteria to higher-risk people limits the generalizability of the results and makes it more difficult to recruit participants into the trial.

The investigator must have reliable estimates of the rate of the primary outcome in people who might be included. These estimates can be based on data from vital statistics, longitudinal observational studies, or rates observed in the untreated group of trials with outcomes similar to those in the planned trial. For example, expected rates of breast cancer in 60-year-old women can be estimated from cancer registry data, from large observational studies such as the Nurses Health Study (1), or from rates of breast cancer observed in large trials of mammographic screening.

Including participants with a high risk of the outcome can decrease the number of subjects needed for the trial. If risk factors for the outcome have been established, then the selection criteria can be designed to include participants who have a minimum estimated risk of the outcome of interest. The Breast Cancer Prevention Trial of tamoxifen for prevention of breast cancer, for example, applied combinations of risk factors for breast cancer in an algorithm designed to include women who had a 5-year risk of breast cancer of at least 1.7% (2). Another way to increase the rate of events is to limit enrollment to people who already have the disease. The Heart and Estrogen/Progestin Replacement Study (HERS) included 2,763 women who already had coronary heart disease (CHD) to test whether estrogen plus progestin reduced the risk of new CHD events (3). This approach was much

more feasible and less costly than a study of women without CHD, which would require about 25,000 participants (4).

Additionally, a trial can be smaller and shorter if it includes people who are likely to have the greatest effect from the treatment. For example, tamoxifen appears to decrease the risk of breast cancer that is estrogen receptor positive but not that of cancer that is estrogen receptor negative. Thus a trial testing the effect of tamoxifen on the risk of breast cancer would be smaller and shorter if participants are at high risk of estrogen receptor positive breast cancer (5,6).

Limiting inclusion to participants at high risk of the disease has two disadvantages. The results of the trial may not be generalizable to lower-risk populations. For example, because the HERS study included only women with CHD, it left uncertainty about whether the findings could be generalized to women without CHD. Furthermore, documenting the basis of the increased risk, such as a history of CHD, may require collecting data or making measurements that make recruitment of participants more complicated, difficult, and expensive.

Although probability samples of general populations confer advantages in observational studies, this type of sampling is generally not feasible and has limited value for most randomized trials. Inclusion of participants with diverse characteristics will increase the confidence that the result of a trial applies broadly. However, setting aside issues of adherence to randomized treatment, it is generally true that results of a trial done in a convenience sample (e.g., women with CHD who respond to advertisements) will be similar to results obtained in probability samples of eligible people (all women with CHD).

Exclusion criteria should be parsimonious because unnecessary exclusions may diminish the generalizability of the results, make it more difficult to recruit the necessary number of participants, and increase the complexity and cost of recruitment. There are five reasons for excluding people from a clinical trial (Table 10.1).

The treatment may be unsafe in people who are susceptible to known or suspected adverse effects of the active treatment. For example, tamoxifen increases the risk of venous thrombosis so that women who have a history of thromboembolic disease should not take the drug and should be excluded. Conversely, the treatment may be known to be so beneficial for some people, such as women with recently diagnosed estrogen receptor positive breast cancer, that it would be unacceptable to assign them to the placebo group. Persons in whom the active treatment is unlikely to be effective should be excluded, as well as those who are unlikely to be adherent to the intervention or unlikely to complete follow-up. It is wise to exclude people who are not likely to contribute a primary outcome to the study (e.g., because they will move during the period of follow-up). Occasionally, practical problems, such as impaired mental status that makes it difficult to follow instructions, justify exclusion. Investigators should carefully weigh potential exclusion criteria that apply to many people (e.g., diabetes or upper age limits) as these have a large impact on the feasibility and costs of recruitment and generalizability of results.

Design an Adequate Sample Size and Plan the Recruitment Accordingly

Trials with too few participants to detect substantial effects are wasteful, unethical, and may produce misleading conclusions (7). Estimating the sample size is one of the most important early parts of planning a trial (Chapters 5 and 6). Recruitment for a trial is usually more difficult than recruitment for an observational study because participants must be willing to be randomly assigned and to take

TABLE 10.1
Reasons for Excluding People from a Clinical Trial

Reason	Example (A trial of tamoxifen vs. placebo to prevent breast cancer)
1. A study treatment would be harmful	
Unacceptable risk of adverse reaction to active treatment	Prior venous thromboembolic event (tamoxifen increases risk of venous thromboembolic events)
Unacceptable risk of assignment to placebo	Recent estrogen receptor–positive breast cancer (treatment with tamoxifen is an effective standard of care)
2. Active treatment cannot or is unlikely to be effective	
Not at risk for the outcome	Bilateral mastectomy
Has a type of disease that is not likely to respond to treatment	Has a breast cancer susceptibility gene that causes estrogen receptor negative cancer
Taking a treatment that is likely to interfere with the intervention	Taking estrogen (which competes with tamoxifen)
3. Unlikely to adhere to the intervention	Poor adherence during run-in
4. Unlikely to complete follow-up	Plans to move before trial ends
	Short life expectancy because of a serious illness
	Unreliable participation in visits before randomization
5. Practical problems with participating in the protocol	Impaired mental state that prevents accurate answers to questions

a blinded therapy. For this reason, the investigator should plan a large, accessible population and enough time and money to get the desired sample size when (as usually happens) the barriers to doing so turn out to be greater than expected.

MEASURING BASELINE VARIABLES

Collect Tracking Information

In addition to the participant's name and address, it is important to record information, such as the names, phone numbers, addresses, and email addresses of two or three friends or relatives who will always know how to reach the participant. It is also valuable to record Social Security numbers or other national I.D. numbers. These can be used to determine the vital status of participants (through the National Death Index) or to help determine the occurrence of key outcomes using health records (e.g., health insurance systems).

Describe the Participants

Investigators should collect enough information (e.g., age, gender, and measurements of the severity of disease) to help others judge the generalizability of the findings. These measurements also provide a means for checking on the comparability of the study groups at baseline; the first table of the final report of a clinical trial typically compares the levels of baseline characteristics in the two study groups. The goal is to make sure that differences in these levels do not

exceed what might be expected from the play of chance, which might suggest a technical error or bias in carrying out the randomization.

Measure Variables that Are Risk Factors for the Outcome or that Can Be Used to Define Subgroups

Particularly in relatively small trials, it is a good idea to measure baseline variables that are likely to be strong predictors of the outcome (smoking habits of the spouse in a trial of a smoking intervention, for example). This allows the investigator to examine these predictors as secondary research questions, and it permits statistical adjustment of the primary randomized comparison to reduce the effects of chance maldistributions of baseline factors between the two study groups, increasing the efficiency of the study. It also allows the investigator to examine whether the intervention has different effects in **subgroups** classified by baseline variables, a relatively uncommon but sometimes important phenomenon termed **effect modification** or **interaction**.

Establish Banks of Materials

Storing sera, DNA, or other biologic specimens at baseline will allow subsequent measurement of biologic predictors of the outcome, and factors such as genotypes, that might identify interactions (subgroups who respond well or poorly to the treatment). Stored specimens can also be a rich resource to study other research questions not directly related to the main outcome.

Measure the Outcome Variable

If possible, it is generally useful to measure the outcome variable at the beginning of the study as well as at the end. In studies that have a dichotomous outcome (the incidence of CHD, for example) it may be important to demonstrate by history and electrocardiogram that the disease is not present at the outset. In studies that have a continuous-outcome variable (a study of the effects of antihypertensive drugs on blood pressure, for example) the best outcome measure is generally the degree of change over the course of the study. This approach controls for differences among the study participants in their initial blood pressure levels and may offer more power than simply comparing blood pressure values at the end of the trial. Similarly, it may also be useful to measure secondary outcomes, and outcomes of planned ancillary studies at baseline.

Be Parsimonious

Having pointed out all these uses for baseline measurements, we should stress that the basic design of a clinical trial does not require that *any* be measured, because randomization eliminates the problem of confounding by factors that are present at the outset. Making a lot of measurements adds expense and complexity. In a randomized trial that has a limited budget, time and money are usually better spent on things that are vital to the integrity of the study, such as the sufficiency of the sample size, the success of randomization and blinding, and the completeness of follow-up. Yusuf et al. have promoted the use of large trials with very few measurements (8).

■ RANDOMIZING

The third step in Fig. 10.1 is to randomly assign the participants to two or more interventions. In the simplest design, one group receives an active treatment and the other receives a placebo.

The random allocation of participants to one or another of the study groups establishes the basis for testing the statistical significance of differences between these groups in the measured outcome. Random assignment provides that age, sex, and other prognostic baseline characteristics that could confound an observed association will be distributed equally, except for chance variation, among the randomized groups. The effects of any maldistributions that do occur as a result of chance are addressed in the statistical tests of the difference in outcome between the randomly assigned groups.

Do a Good Job of Random Assignment

Because randomization is the cornerstone of a clinical trial, it is important that it be done correctly. The two most important features are the following: (a) The procedure truly allocates treatments randomly and (b) the assignments are tamperproof so that neither intentional nor unintentional factors can influence the randomization.

Ordinarily, the participant completes the baseline examinations, is found eligible for inclusion, and gives consent to enter the study before randomization. He is then randomly assigned by applying a previously established algorithm to a set of random numbers, which are typically computer-generated. For example, if the design calls for an equal probability of assignment to each of three study groups, the algorithm could specify using the random numbers 1, 2, and 3 for assignment to the three study groups. (Other numbers are ignored.)

Once a list of the random order of assignment to study groups is generated, it must be applied to participants as they enter the trial. It is essential to design the random assignment procedure so that members of the research team who have any contact with the study participants cannot influence the allocation. Random treatment assignments can be placed in advance in a set of sealed envelopes by someone who will not be involved in opening the envelopes. Each envelope must be numbered (so that all can be accounted for at the end of the study), opaque (to prevent transillumination by a strong light), and otherwise tamperproof. When a participant is randomized, his name and the number of the next envelope are first recorded; then the envelope is opened. For small studies a table of random numbers (Appendix 3.1) can be used to order the envelopes.

An effective but costly procedure for ensuring that randomization is tamperproof is to set up a separate randomization facility that the trial staff contact by telephone, fax, or email when an eligible participant is ready to be randomized. The staff member provides the name and study number of the new participant. This information is recorded and the treatment group is then randomly assigned. Treatment can also be randomly assigned by computer programs at the research site as long as these programs are tamperproof.

Rigorous precautions to prevent tampering with randomization are needed because investigators sometimes find themselves under pressure to influence the randomization process (e.g., for an individual who seems particularly suitable for an active treatment group in a placebo-controlled trial).

Consider Special Randomization Techniques

In general, the preferred approach is to randomize equal numbers of participants to each group, but unequal allocation has been used in studies that have three or more groups, one serving as a control for each of the others (9). If no formal comparisons among the active treatment groups are planned, the larger number of comparisons that involve the control group makes the precision of its outcome

measure especially important and the investigator can consider assigning a larger proportion of participants to the control group. However, there is no clear way to pick the best proportions to use, and disproportionate randomization in this and other situations can complicate the process of obtaining informed consent. Because the advantages are marginal (the effect of even a 2:1 disproportion on power is surprisingly modest [10]), the best decision is usually to assign equal numbers to each group.

Trials of small to moderate size will have a small gain in power if special randomization procedures are used to balance the study groups in the numbers of participants they contain (blocked randomization) and in the distribution of baseline variables known to predict the outcome (stratified blocked randomization).

Blocked randomization is a commonly used technique to ensure that the number of participants is equally distributed among the study groups. Randomization is done in "blocks" of predetermined size. For example, if the block size is six, randomization proceeds normally within each block until the third person is randomized to one group, after which participants are automatically assigned to the other group until the block of six is completed. This means that in a study of 30 participants exactly 15 will be assigned to each group, and in a study of 33 participants, the disproportion could be no greater than 18:15. Blocked randomization is less suitable for nonblinded studies because the treatment assignment of the participants at the end of each block could be predicted and manipulated. This problem can be minimized by varying the size of the blocks randomly (ranging, for example, from four to eight) according to a schedule that is not known to the investigator.

Stratified blocked randomization ensures that an important predictor of the outcome is more evenly distributed between the study groups than chance alone would dictate. In a trial of the effect of a drug to prevent fractures, having a vertebral fracture is such a strong predictor of outcome and response to treatment that it may be best to ensure that very similar numbers of people who have vertebral fractures are assigned to each group. This can be achieved by dividing the study cohort at baseline into participants with or without vertebral fractures, and then carrying out a blocked randomization within each of these two "strata." Stratified blocked randomization can slightly enhance the power of a small trial by reducing the variation in outcome due to chance disproportions in important baseline variables. Stratified blocked randomization is of little benefit in large trials (more than 1,000 participants) because chance assignment ensures nearly even distribution of baseline variables. An important limitation of stratified randomization is the small number of baseline variables, not more than two or three, that can be balanced by this technique.

■ APPLYING THE INTERVENTIONS

In a clinical trial the investigator compares the outcome in groups of participants that receive different interventions. Between-group designs always include a group that receives a treatment to be tested, and a control group that receives either no active treatment (placebo) or a standard comparison treatment. Sometimes there are also additional experimental groups that receive other treatments. There are a number of trade-offs to consider in choosing and applying these interventions.

■ TABLE 10.2
Randomization Eliminates Confounding by Baseline Variables and Blinding Eliminates Confounding by Co-Interventions

Explanation for Association	Strategy to Rule Out Rival Explanation
1. Chance	Same as in observational studies
2. Bias	Same as in observational studies
3. Effect-Cause	(Not a possible explanation in an experiment)
4. Confounding — Prerandomization confounding variables	**Randomization**
4. Confounding — Postrandomization confounding variables (unintended interventions)	**Blinding**
5. Cause-Effect	

Importance of Blinding

Whenever possible, the investigator should design the interventions in such a fashion that neither the study participants nor anybody who has contact with them has any knowledge of the study group assignment. In a randomized trial, *blinding is as important as randomization:* It protects the trial from treatment differences between the groups other than the randomized treatments and from biased assessment of outcomes.

Randomization only eliminates the influence of confounding variables that are present at the time of randomization; it does not eliminate confounding that occurs during follow-up (Table 10.2). In an unblinded study the investigator may give extra attention to participants he knows are receiving the active drug, and this "**co-intervention**" may be the actual cause of any difference in outcome that is observed between the groups. For example, in an unblinded trial of the effect of exercise to prevent myocardial infarction, the investigator's eagerness to find a benefit might lead him to suggest that participants in the exercise group adopt a low-fat diet and stop smoking. Co-interventions can also affect the control group if, for example, participants who discover that they are receiving placebo seek out other treatments that affect the difference in outcome between groups.

Sometimes interventions cannot be blinded. In this case, the investigator should limit and standardize other potential treatments as much as possible. For example, an investigator testing the effect of exercise for reduction in hot flashes could specify a precise regimen of exercise sessions in the treatment group. To minimize other differences in exercise, he could instruct both exercise and control participants to refrain from starting new recreational activities or hormone treatment until the trial has ended.

The second important value of blinding is to prevent **biased assessment of outcome.** In an unblinded trial, the investigator may be tempted to look more carefully for outcomes in the untreated group or to diagnose the outcome more frequently. For example, in an unblinded trial of estrogen therapy, the investigators may be more likely to ask women in the active treatment group about pain or swelling in the calf and to order ultrasound or other tests to make the diagnosis of deep vein thrombosis. Blinded assessment of outcome may not be important if the outcome of the trial is death, about which there is no uncertainty or opportunity for biased assessment. Most other outcomes, such as cause-specific death, disease diagnosis, physical measurements, questionnaire scales, and self-reported conditions, are susceptible to biased ascertainment. Blinded assessment of outcomes is most important in trials where the outcome is "soft," such as those based on participant self-report or investigator opinion. In a trial of the effect of exercise on depression, participants who receive the exercise program may report less depression due to a placebo effect, and the investigator (who wants the intervention to be effective) may be more likely to judge that participants who were assigned to exercise are less depressed. Especially when the intervention cannot be blinded, outcomes should be "hard" (i.e., based on measurements that are resistant to bias). Measurement of depression, for example, is less susceptible to bias if change in a standardized depression scale is used, rather than investigator judgment.

Blinding is more difficult to carry out successfully than randomization. Even when an intervention such as a drug can be blinded, the logistic problems can be substantial. There is the need to get the manufacturer or pharmacy to prepare the identical capsules and to develop foolproof systems for labeling and dispensing. Well before a trial starts, investigators should inspect, taste, and weigh samples of the active drug and placebo and their containers. (Investigators sometimes discover differences in the weight of bottles and taste or odor of pills only after randomization has begun—too late to change the preparations.) In addition, it may be necessary to develop a 24-hour mechanism for unblinding in the event that a participant becomes acutely ill and his personal physician needs to know what drug he is taking. Pharmacies will sometimes help to design and provide this service.

The other major difficulty in designing the system for blinding is ensuring that neither the participants nor the research team will be able to discern the effects of treatment assignment. Telltale effects of the drugs on physical symptoms or laboratory values (such as the effect of diuretics on serum potassium) may require setting up a system in which these results are reviewed by someone not involved in follow-up contacts with the participants. In the HERS trial, even though the active hormones and the placebo were provided in identical capsules, treatment with hormones caused vaginal bleeding in a substantial proportion of postmenopausal women assigned to active treatment. To address this potential unblinding, participants were instructed to report any bleeding to the gynecologic staff who did not interact with other trial staff and had no role in the ascertainment of the main outcome—coronary disease events (11).

After the study is over, it is a good idea to assess whether the participants and investigators were unblinded by asking them to guess the treatment assignments; if a higher than expected proportion guesses correctly, the published discussion of the findings can include an assessment of the potential biases the partial unblinding may have caused.

Choice of Intervention

Investigators should consider several issues as they design their interventions, including the balance of effectiveness and safety, the feasibility of blinding, whether to treat with one or a combination of interventions, and generalizability to the way the treatment will be used in practice (e.g., whether to use a fixed treatment dose or titrate). If important decisions are uncertain, such as which dose best balances effectiveness and safety, it may be best to postpone a trial until pilot studies have been completed to help resolve the issue. However, even after gathering a lot of data about the alternatives, the best approach is often uncertain. (If the best approach were clear, there would be no need for the study.) Sometimes an investigator may hedge his bets by including two or more treatments in order to test each of several promising interventions. For example, at the time the Multiple Outcomes of Raloxifene Trial was designed, it was not clear which dose of raloxifene, 60 or 120 mg, was best, so the trial tested two doses of raloxifene for preventing fractures (9). This is sometimes a reasonable strategy, but it has its costs: a larger and more expensive trial, and the complexity of dealing with multiple hypotheses (Chapter 5). Choosing the best treatment can be especially difficult in studies that involve years of follow-up because a treatment that reflects current practice at the outset of the study may have become outmoded by the end, transforming a pragmatic test into an academic exercise.

The best balance between effectiveness and safety depends on the condition being treated. On the one hand, effectiveness is generally the paramount consideration in designing interventions to treat illnesses that cause severe symptoms and a high risk of death. Thus it may be best to choose the "highest tolerable dose." On the other hand, safety should be the primary criterion for designing interventions to reduce the risk of less severe and nonfatal conditions among healthy people: Few people will benefit from the treatment by avoiding the condition, all will be at risk of any adverse effects of the drug. In this case, it is generally best to choose the "lowest effective dose."

The development and testing of new drugs generally involves a series of relatively small randomized trials of a range of doses or combinations of treatments (called Phase II trials by the Food and Drug Administration; see Chapter 11). At this stage, it is wise to test the widest possible range of doses and regimens to guarantee that an effective and ineffective dose are included. The outcomes are often physiologic measurements on continuous scales (such as change in cholesterol level) because it is not feasible to study the effect of multiple doses on the real outcome of interest (myocardial infarction). Investigators should carefully consider the evidence that surrogate markers are likely to predict accurately the effect of treatment on the outcome of interest. When the outcome, such as myocardial infarction, depends on several processes (levels of HDL and LDL cholesterol, blood coagulability, platelet adhesiveness, endothelial cell function), the investigator should measure several surrogate markers.

Because of their limited size, it is usually not possible for Phase II studies to assess the safety of a new drug adequately. The effect of drugs on laboratory tests, such as liver function tests, is tested because these tests are easy and inexpensive. Unless the treatment has a common and severe toxicity, it may be impossible to assess the safety of a treatment until it is tested in larger trials that have adequate power to detect uncommon but important adverse effects. This reinforces the wisdom of choosing the lowest dose that appears to have adequate effectiveness.

When a choice is possible, investigators should clearly prefer interventions that can be blinded over alternatives that cannot. Creative approaches can help blind, or partially blind, many outcomes. For example, in a trial of the effect of acupuncture on pain from dysmenorrhea, the treated group can have needles placed in points thought to relieve pain, and the control group can have needles placed in other locations.

Trials to test single interventions are generally much easier to plan and implement than those testing combinations of treatments. However, many medical conditions, such as HIV infection, are treated with combinations of drugs. The most important disadvantage of testing combinations of treatments is that the result cannot provide clear conclusions about any one of the interventions. In the HERS trial, for example, postmenopausal women were treated with estrogen plus progestin therapy or placebo. The intervention did not result in a reduced risk of coronary events, but it was unclear whether estrogen alone may have had a more beneficial effect. In general, it is preferable to design trials that have only one major difference between any two study groups.

The investigator should consider how well the intervention can be incorporated in practice. Thus simple interventions are generally better than complicated ones (patients are more likely to take a pill once a day than two or three times). Complicated interventions, such as multifaceted counseling of patients about changing their behavior, may not be feasible to incorporate in general practice because they require rare expertise or are too time-consuming or costly. Such interventions are less likely to have clinical impact, even if a trial proves that they are effective.

Some treatments are generally given in doses that vary from patient to patient. In these instances, it may be best to design an intervention so that the active drug is titrated to achieve a clinical outcome such as reduction in the hepatitis C viral load. To maintain blinding, corresponding changes should be made (by someone blinded to treatment group) in the "dose" of the placebo.

Choice of Control

The best control group receives no active treatment in a way that can be blinded, which for medications means receiving a placebo that is identical to active treatment. This strategy compensates for any placebo effect of the active intervention (i.e., through suggestion and other nonpharmacologic mechanisms) so that any outcome difference between study groups can be ascribed to a biologic effect.

Often, however, it is not possible or desirable to withhold all treatment. For example, in trials of statin drugs to reduce the risk of myocardial infarction in persons with known CHD, the investigators cannot ethically prohibit or discourage participants from taking medical treatments that are indicated for persons with known CHD, including aspirin and beta-blockers. It is important to consider such "co-interventions" for two reasons. First, if participants use medications or therapies (other than the study drug) that reduce the risk of developing the outcome of interest, the reduction in power means that the sample size will need to be larger or the trial longer. Second, the trial protocol must include plans to obtain data to allow statistical adjustment for differences between the groups in the rate of use of such "co-interventions" during the trial. However, adjusting for such postrandomization differences violates the intention-to-treat principle and should be viewed as a secondary or explanatory analysis (Chapter 11). Alternatively, the investigators may give a standard

treatment, such as aspirin, to all participants in the trial. Although this approach reduces power, it minimizes the potential for differences in co-interventions between the groups and tests whether the new intervention improves outcome when given in addition to standard care.

For some medical conditions, there is already a standard treatment. In this case, new treatments should be evaluated by comparison with those already proven effective. These are known as **equivalence trials**. Ideally, the new treatment should have advantages—lower cost, less frequent administration, or greater safety—so that finding no difference in outcome would lead to the conclusion that one agent is superior. If the new treatment has no advantages, then demonstrating that it has an effectiveness that is similar to that of an older one may be valuable to the manufacturer of the new treatment, but the effort produces no advance in clinical care or public health.

In trials with active treatment as the control group, the challenge is to show convincingly that, if no significant difference is seen, the two treatments are equivalent. Statistical approaches involve testing whether the observed difference in outcomes between the two groups lies within some specified range (12). This may require larger sample sizes than trials that have placebo controls because the difference to be detected may be smaller.

As noted in Chapter 6, a basic problem with equivalence studies is that because the goal is to accept the null hypothesis rather than reject it, the normal strategies that reduce the likelihood of Type I errors do not have their usual effect in safeguarding the conclusions. The failure to find a statistically significant difference between active treatments in equivalence trials can result from designing a randomized trial that has too few participants, too few outcomes, or imprecise measurements of outcome.

■ SUMMARY

1. The criteria for **selecting study participants** should provide subjects at **high risk** of the outcome (if dichotomous), likely to **benefit and not be harmed** by treatment, easy to recruit, and likely to **adhere to treatment** and **follow-up** protocols.
2. **Baseline variables** should be measured parsimoniously (setting aside banks of serum, genetic material, and so on, for later analysis) to **describe** the participants, to **measure risk factors** and baseline values of the outcome, and to allow examination of **interactions** in which the intervention has different effects in different subgroups.
3. **Randomization,** which eliminates bias due to baseline **confounding** variables, should be tamperproof; in small trials **stratified blocked randomization** can reduce the impact of chance maldistributions of key predictors.
4. **Blinding the intervention** is as important as randomization and serves to control **co-intervention** and **ascertainment biases.**
5. The **choice of intervention** is a difficult decision that balances **effectiveness** and **safety;** other considerations include the need for **relevance** to clinical practice and the scientific utility of **single interventions** in preference to combinations.
6. Whenever possible, trials should include a comparison with a **placebo control;** the danger of comparing active treatments in an **equivalence trial** is the reversal of the normal safeguards of testing the null hypothesis.

References

1. Colditz GA, Hankinson SE, Hunter DJ, et al. The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *N Engl J Med* 1995;332:1589–93.
2. Fisher B, Costantino JP, Wickerham DL, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. *J Natl Cancer Inst* 1998;90:1371–88.
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. Heart and Estrogen/Progestin Replacement Study (HERS) Research Group. *JAMA* 1998; 280:605–13.
4. 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.
5. Fisher B, Dignam J, Wolmark N, et al. Tamoxifen in treatment of intraductal breast cancer: National Surgical Adjuvant Breast and Bowel Project B-24 randomised controlled trial. *Lancet* 1999;353:1993–2000.
6. Cummings SR, Eckert S, Krueger KA, et al. The effect of raloxifene on risk of breast cancer in postmenopausal women: results from the MORE randomized trial. Multiple outcomes of raloxifene evaluation. *JAMA* 1999;281:2189–97.
7. Freiman JA, Chalmers TC, Smith H, Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med* 1978;299:690–4.
8. Yusuf S, Collins R, Peto R. Why do we need some large, simple randomized trials? *Stat Med* 1984;3:409–22.
9. Ettinger B, Black DM, Mitlak BH, et al. Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene: results from a 3-year randomized clinical trial. Multiple outcomes of raloxifene evaluation (MORE) investigators. *JAMA* 1999;282:637–45.
10. Friedman LM, DeMets DL, Furberg C. *Fundamentals of clinical trials*, 3rd ed. St. Louis: Mosby-Year Book, 1996.
11. Grady D, Applegate W, Bush T, et al. Heart and Estrogen/Progestin Replacement Study (HERS): design, methods, and baseline characteristics. *Control Clin Trials* 1998;19:314–35.
12. Hauck WW, Anderson S. A proposal for interpreting and reporting negative studies. *Stat Med* 1986;5:203–9.