CHAPTER 11

Alternative Clinical Trial Designs and Implementation Issues

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

In the last chapter, we discussed the classic randomized, blinded, parallel group trial: how to select and blind the intervention and control conditions, randomly assign the interventions, choose outcomes, deal with adverse events, select participants, and measure baseline and outcome variables.

In this chapter, we describe alternative **randomized and non-randomized** between-group **trial designs**, as well as **within-group** designs, **cross-over** studies, and **pilot studies**. We then address the **conduct of clinical trials**, including **adherence to the intervention and follow-up**, and **ascertaining and adjudicating outcomes**. We conclude with a discussion of statistical issues such as **interim monitoring** for stopping the trial early, **intention to treat** and **per-protocol** analyses, and the use of **subgroup analysis** to discover effect modification.

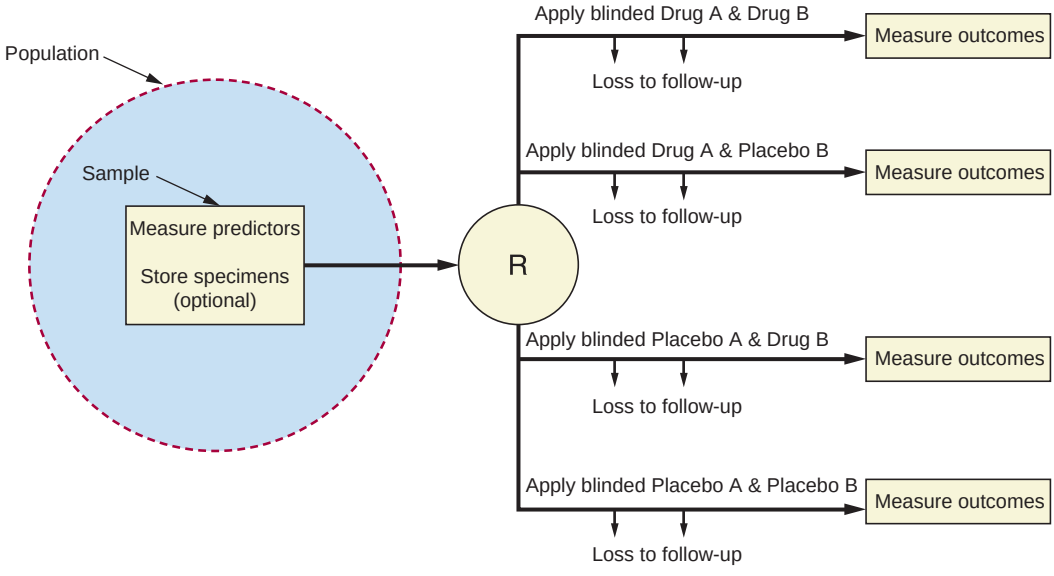
ALTERNATIVE RANDOMIZED DESIGNS

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

Factorial Design

The **factorial design** aims to answer two (or more) separate research questions in a single trial (Figure 11.1). A good example is the Women's Health Study, which was designed to test the effect of low-dose aspirin and of vitamin E on the risk for cardiovascular events among healthy women (1). The participants were randomly assigned to four groups, and two hypotheses were tested by comparing two halves of the study cohort. First, the rate of cardiovascular events in women on aspirin was compared with women on aspirin placebo (disregarding the fact that half of each of these groups received vitamin E); then the rate of cardiovascular events in those on vitamin E was compared with all those on vitamin E placebo (now disregarding the fact that half of each of these groups received aspirin). The investigators have two complete trials for the price of one.

A limitation is the possibility of **effect modification** (interaction): if the effect of aspirin on risk for cardiovascular disease is different in women treated with vitamin E than in those not treated with vitamin E, effect modification is present and the effect of aspirin would have to be calculated separately in these two groups. This would reduce the power of these comparisons, because only half of the participants would be included in each analysis. Factorial designs can actually be used to *study* effect modification, but trials designed for this purpose are more complicated and difficult to implement, larger sample sizes are required, and the results can be hard to interpret. Other limitations of the factorial design are that the same study population must be appropriate for each intervention, multiple treatments may interfere with recruitment and adherence, and analyses are more complex. That said, the factorial design can be very **efficient**. For example, the Women's Health Initiative randomized trial was able to test the effect of three



■ **FIGURE 11.1** In a factorial randomized trial, the steps are to:

- Select a sample of participants from a population suitable for receiving the intervention.
- Measure the predictor variables and (if appropriate) the baseline level of the outcome variable.
- Consider the option of storing serum, images, and so on, for later analysis.
- Randomly assign two (or more) active interventions and their controls to four (or more) groups.
- Follow the cohorts over time, minimizing loss to follow-up and assessing adherence to the intervention and control conditions.
- Measure the outcome variables.
- Analyze the results, first comparing the two intervention A groups (combined) to the combined placebo A groups and then comparing the two intervention B groups (combined) to the combined placebo B groups.

interventions (postmenopausal hormone therapy, low-fat diet, and calcium plus vitamin D) on a number of outcomes (2).

Cluster Randomization

Cluster randomization requires that the investigator randomly assign naturally occurring groups or clusters of participants to the interventions, rather than 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 compared to control teams (3). 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 only one member of a family. When participants in a natural group are randomized individually, those who receive the intervention are likely to discuss or share the intervention with family members, colleagues, team members, or acquaintances who have been assigned to the control group.

In the cluster randomization design, the units of randomization and analysis are groups, not individuals. Therefore, the effective sample size is smaller than the number of individual participants and power is diminished. The effective sample size depends on the correlation of the effect of the intervention among participants in the clusters and is somewhere between the

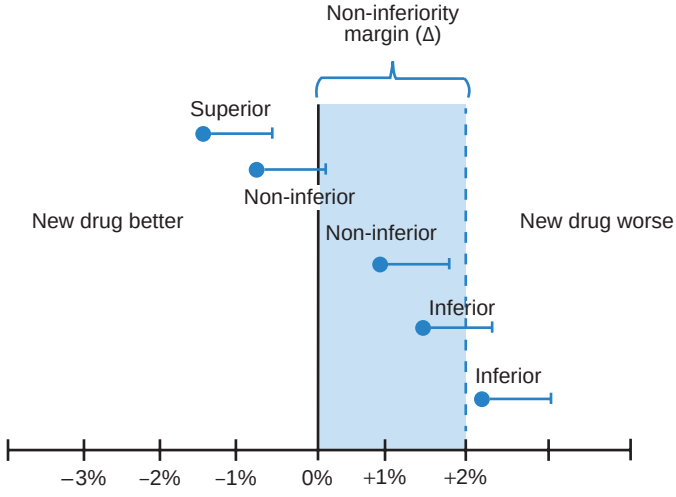
number of clusters and the number of participants (4). Other drawbacks are that sample size estimation and data analysis are more complicated in cluster randomization designs than for individual randomization (4).

Active Control Trials: Equivalence and Non-Inferiority

An **active control trial** is one in which the control group receives an active treatment. This design may be optimal when there is a known effective treatment or "standard of care" for a condition. This type of trial is sometimes called a **comparative effectiveness trial** because two treatments are compared.

In some cases, the aim of an active control trial is to show that a new treatment is **superior** to an established treatment. In this situation, the design and methods are similar to a placebo-controlled trial. In most cases, however, investigators want to establish that a new therapy that has some advantages over an established therapy (easier to use, less invasive, safer) has *similar* efficacy. In this case, an **equivalence** or **non-inferiority** trial is more appropriate.

The **statistical methods** for equivalence or non-inferiority trials are different than for trials designed to show that one treatment is better than another. In a trial designed to show that a treatment is superior, the standard analysis uses tests of statistical significance to accept or reject the null hypothesis that there is no difference between groups. In a trial designed to show that a new treatment is equivalent to the standard treatment, on the other hand, the ideal goal would be to *accept* the null hypothesis of no difference. But proving that there is *no* difference between treatments (not even a tiny one) would require an infinite sample size. So the practical solution is to design the sample size and analysis plan using a confidence interval (CI) approach—considering where the CI for the effect of the new treatment compared to the standard treatment lies with respect to a prespecified delta ("Δ"), the unacceptable difference in efficacy between the two treatments (5, 6). Equivalence or non-inferiority is considered established at the level of significance specified by the CI if the CI around the difference in efficacy of the new compared to the established treatment does not include Δ (Figure 11.2). This is a



■ **FIGURE 11.2** Possible outcomes in a non-inferiority trial comparing a new drug to warfarin as treatment to reduce stroke risk among patients with atrial fibrillation, with the non-inferiority margin (delta) set at +2%. The one-sided 95% confidence intervals around the difference in stroke rate between warfarin and the new drug are shown illustrating the outcomes of superiority, inferiority, and non-inferiority.

two-tailed consideration in the case of an equivalence trial (i.e., the new treatment is neither worse *nor* better than the standard treatment). However, it is uncommon for investigators to be interested in whether a new treatment is *both* no better *and* no worse than an established treatment. Most often, investigators are especially interested in showing that a new treatment with other advantages is not inferior to the standard treatment. The one-tailed nature of the non-inferiority trial design also has the advantage of permitting either a smaller sample size or a smaller alpha; the latter is usually preferred (e.g., 0.025 rather than 0.05), to be conservative.

One of the most difficult issues in designing a non-inferiority trial is establishing the **non-inferiority margin (Δ)**—the loss of efficacy of the new treatment that would be unacceptable (7). This decision is based on both statistical and clinical considerations of the potential efficacy and advantages of the new treatment, and requires expert judgment (8) (see Appendix 11A for an example of how this is done). Non-inferiority trials generally need to be larger than placebo-controlled trials because the acceptable difference between the new and established treatment is usually smaller than the expected difference between a new treatment and placebo.

It is important to note that non-inferiority may not mean that both the established and new treatments are effective—they could be equivalently ineffective or harmful. To ensure that a new treatment evaluated in a non-inferiority trial is more effective than placebo, there should be strong prior evidence supporting the efficacy of the established treatment. This also means that the design of the non-inferiority trial should be as similar as possible to trials that have established the efficacy of the standard treatment, including selection criteria, dose of the established treatment, adherence to the standard treatment, length of follow-up, loss to follow-up, and so on (6, 7). Any problem that reduces the efficacy of the standard treatment (enrolling participants unlikely to benefit, non-adherence to treatment, loss to follow-up) will make it more likely that the new therapy will be found to be non-inferior—simply because the efficacy of the standard treatment has been reduced. A new, less effective treatment may appear to be non-inferior when, in reality, the findings represent a poorly done study.

In summary, non-inferiority and equivalence trials are particularly worthwhile if a new treatment has important advantages such as lower cost, ease of use, or safety. It is difficult to justify large trials to test a new "me-too" drug with none of these advantages. Importantly, non-inferiority and equivalence trials can produce the misleading conclusion that two treatments are equivalent if the trial is poorly conducted.

Adaptive Designs

Clinical trials are generally conducted according to a protocol that does not change during the conduct of the study. However, for some types of treatments and conditions, it is possible to monitor results from the trial as it progresses and **change the design** of the trial **based on interim analyses** of the results (9). For example, consider a trial of several doses of a new treatment for non-ulcer dyspepsia. The initial design may plan to enroll 50 participants to a placebo group and 50 to each of three doses for 12 weeks of treatment over an enrollment period lasting 1 year. Review of the results after the first 10 participants in each group have completed 4 weeks of treatment might reveal that there is a trend toward relief of dyspepsia only in the highest dose group. It may be more efficient to stop assigning participants to the two lower doses and continue randomizing only to the highest dose and the placebo. Other facets of a trial that could be changed based on interim results include increasing or decreasing the **sample size** or **duration** of the trial if interim results indicate that the effect size or rate of outcomes differ from the original assumptions.

Adaptive designs are feasible only for treatments that produce outcomes that are measured and analyzed early enough in the course of the trial to make design changes in the later stages of the trial possible. To prevent bias, rules for how the design may be changed should be established before the trial begins, and the interim analyses and consideration of change in design should be done by an independent data and safety monitoring board that reviews unblinded

data. Multiple interim analyses will increase the probability of finding a favorable result that is due to chance variation, and the increased chance of a type I error must be considered in the analysis of the results.

In addition to being more complex to conduct and analyze, adaptive designs require that informed consent include the range of possible changes in the study design, and it is difficult to estimate the cost of an adaptive trial and the specific resources needed to complete it. Despite these precautions and limitations, adaptive designs are efficient and may be valuable, especially during the development of a new treatment; they can allow earlier identification of the best dose and duration of treatment, and ensure that a high proportion of participants receive the optimal treatment.

■ NONRANDOMIZED DESIGNS

Nonrandomized Between-Group Designs

Trials that compare groups that have not been randomized are far less effective than randomized trials in controlling for confounding variables. For example in a trial of the effects of coronary artery bypass surgery compared to percutaneous angioplasty, if clinicians are allowed to decide which patients undergo the procedures rather than using random allocation, patients chosen for surgery are likely to be different than those chosen for angioplasty. 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. When the findings of randomized and nonrandomized studies of the same research question are compared, the apparent benefits of intervention are often greater in the nonrandomized studies, even after adjusting statistically for differences in baseline variables (10). The problem of confounding in nonrandomized clinical studies can be serious and not fully removed by statistical adjustment (11).

Sometimes participants are allocated to study groups by a **pseudorandom** mechanism. For example, every participant with an even hospital record number may be assigned to the treatment group. Such designs may offer logistic advantages, but the predictability of the study group assignment permits the investigator or the study staff to tamper with it by manipulating the sequence or eligibility of new participants.

Participants are sometimes assigned to study groups by the investigator according to certain specific criteria. For example, patients with diabetes 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 with this design is that those willing to take four injections per day might differ from those who are unwilling (for example, being 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 than randomization because they allow the participant or clinician to choose the intervention. In fact, studies are only ethical if they have a reasonable likelihood of producing the correct answer to the research question, and randomized studies are more likely to lead to a conclusive and correct result than nonrandomized designs. Moreover, the ethical basis for any trial is the uncertainty as to whether the intervention will be beneficial or harmful. This uncertainty, termed **equipoise**, means that an evidence-based choice of interventions is not possible and this justifies random assignment.

Within-Group Designs

Designs that do not include a separate control group can be useful options for some types of questions. In a **time series design**, measurements are made before and after each participant receives the intervention (Figure 11.3). Therefore, each participant serves as his own control to evaluate the effect of treatment. This means that individual characteristics such as age, sex,

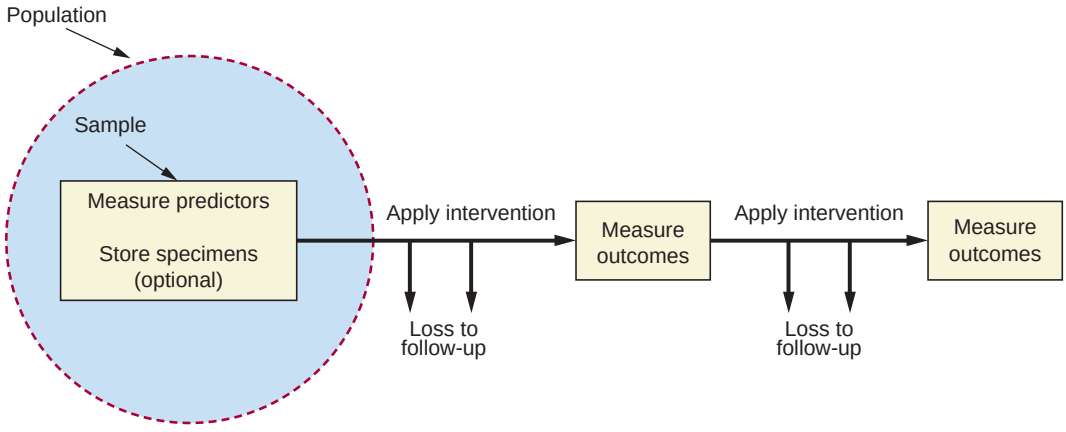


FIGURE 11.3 In a time series trial, the steps are to:

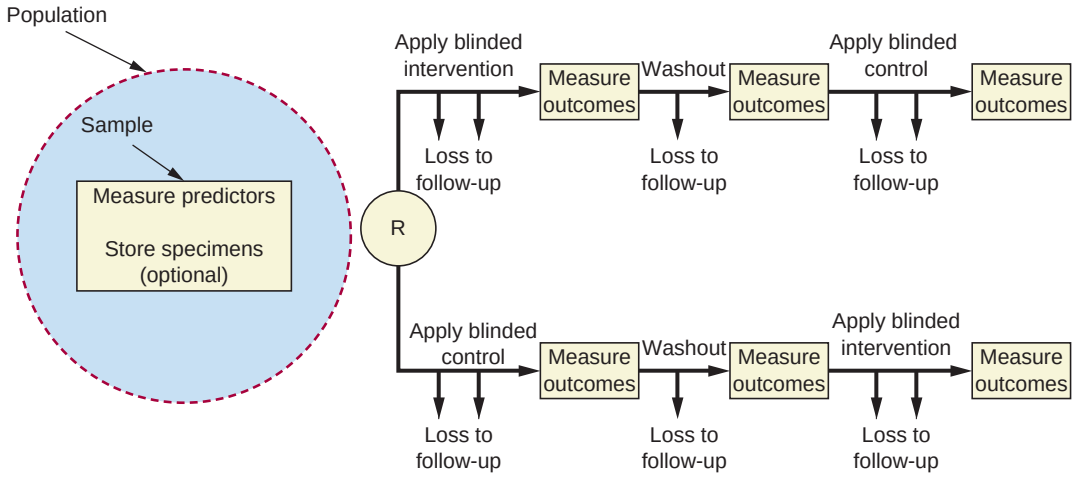
- Select a sample of participants from a population suitable for receiving the intervention.
- Measure the predictor variables and (if appropriate) the baseline level of the outcome variable.
- Consider the option of storing serum, images, and so on for later analysis.
- Apply the intervention to the whole cohort.
- Follow the cohort over time, minimizing loss to follow-up and assessing adherence to the intervention.
- Measure the outcome variables.
- Remove the intervention, continue the follow-up and measure the outcome variable again, then re-initiate the intervention, and so on.

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 flu season ended before follow-up was completed). Within-group designs sometimes use a strategy of repeatedly starting and stopping the treatment. If repeated onset and offset of the intervention produces corresponding patterns in the outcome, this is 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 design has a clinical application in "**N-of-one**" **trials** in which an individual patient can alternate between active and inactive versions of a drug (using identical-appearing placebo prepared by the local pharmacy) to detect his particular response to the treatment (12).

Crossover Designs

The **crossover design** has features of both within- and between-group designs (Figure 11.4). Half of the participants are randomly assigned to start with the control period and then switch to active treatment; the other half begins with the active treatment and then switches to control. This approach permits between-group, as well as within-group, analyses. The advantages are substantial: it minimizes the potential for confounding because each participant serves as his own control and the paired analysis 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, the added expense required to measure the outcome at the beginning and end of each crossover period, and the added complexity of analysis and interpretation created by



■ **FIGURE 11.4** In a crossover randomized trial, the steps are to:

- Select a sample of participants from a population suitable for receiving the intervention.
- Measure the predictor variables and (if appropriate) the baseline level of the outcome variable.
- Randomly assign the blinded intervention and control condition.
- Follow the cohorts over time, minimizing loss to follow-up and assessing compliance with the interventions and control conditions.
- Measure the outcome variables.
- Discontinue the intervention and control condition and provide a washout period to reduce carryover effect, if appropriate.
- Apply intervention to former control group and control condition to former intervention group and measure outcomes after following cohorts over time.

potential **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 not returning 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 between treatments with the hope that the outcome variable will return to baseline before starting the next intervention, but it is difficult to know whether all carryover effects have been eliminated. In general, crossover studies are a good choice when the number of study participants is limited and the outcome responds rapidly and reversibly to an intervention.

A variation on the crossover design may be appropriate when the intervention to be studied cannot be blinded and the intervention is believed by participants to be much more desirable than the control (such as a new noninvasive procedure). In this situation, where it may be very difficult to find eligible participants who are willing to be randomized, an excellent approach may be randomization to immediate intervention versus a **wait-list** (delayed) **control**. Another situation in which a wait-list control may be appropriate is when a community, school, government, or similar entity has decided that all members of a group should receive an intervention, despite limited evidence of efficacy. In this situation, randomization to not receive the intervention may be considered unethical, while randomization to delayed intervention may be acceptable.

The wait-list design provides an opportunity for a randomized comparison between the immediate intervention and wait-list control groups. In addition, the two intervention periods (immediate intervention in one group and delayed intervention in the other) can be pooled to increase power for a within-group comparison before and after the intervention. For example, in a trial in which women with symptomatic fibroids are randomized to a new treatment that is less invasive than hysterectomy (uterine artery embolization) versus wait-list,

the wait-list control would receive no treatment during the initial period; then be offered uterine artery embolization at the beginning of the next period. Subsequently, within-group measurements of changes in fibroid symptom score can be pooled among all of the participants who received the intervention.

This design has the advantage of making enrollment much more feasible in a trial where the intervention is highly desirable, and of allowing a randomized comparison in situations where all eligible participants will eventually receive an intervention. However, the outcome must occur in a short period of time (or the wait period becomes prohibitively long). In addition, providing the intervention to the control group at the end of the trial prolongs the length of follow-up and can be expensive.

Trials for Regulatory Approval of New Interventions

Many trials are done to test the effectiveness and safety of new treatments that might be considered for approval for marketing by the U.S. Food and Drug Administration (FDA) or another regulatory body. 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 design and conduct of these trials is generally the same as for other trials, but regulatory requirements must be considered.

The FDA publishes general and specific guidelines on how such trials should be conducted (search for "FDA" on the Web). It would be wise for investigators and staff conducting trials with the goal of obtaining FDA approval of a new medication or device to seek specific training on general guidelines, called **Good Clinical Practice** (Chapter 17). In addition, the FDA provides specific guidelines for studies of certain outcomes. For example, studies designed to obtain FDA approval of treatments for hot flashes in menopausal women must currently include participants with at least seven hot flashes per day or 50 per week. FDA guidelines are regularly updated and similar guidelines are available from international regulatory agencies.

Trials for regulatory approval of new treatments are generally described by phase. This system refers to an orderly progression in the testing of a new treatment, from experiments in animals, human cell cultures or tissues (**preclinical**) and initial unblinded, uncontrolled treatment of a few human volunteers to test safety (**phase I**), to small randomized or time series trials that test the effect of a range of doses on adverse effects and biomarkers or clinical outcomes (**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 11.1). The FDA usually defines the endpoints for phase III trials that are required to obtain approval to market the new drug. **Phase IV** refers to large studies that may be randomized trials, but are often large observational studies that are conducted after a drug is approved. These studies are often designed to assess the rate of serious

TABLE 11.1 STAGES IN TESTING NEW THERAPIES

Preclinical	Studies in cell cultures, tissues, and animals
Phase I	Unblinded, uncontrolled studies in a few volunteers to test safety
Phase II	Relatively small randomized or time series trials to test tolerability and different intensity or dose of the intervention on biomarkers or clinical outcomes
Phase III	Relatively large randomized blinded trials to test conclusively the effect of the therapy on clinical outcomes and adverse events
Phase IV	Large trials or observational studies conducted after the therapy has been approved by the FDA to assess the rate of uncommon serious side effects and evaluate additional therapeutic uses

side effects when the drug is used in large populations or to test additional uses of the drug that might be approved by the FDA. Sometimes, phase IV studies do not have a clear scientific goal, but are performed to introduce physicians and patients to new drugs.

Pilot Studies

Designing and conducting a successful clinical trial requires extensive information on the type, dose, and duration of the intervention; the likely effect of the intervention on the outcome; potential adverse effects; the feasibility of recruiting, randomizing, and maintaining participants in the trial; and likely costs. Often, the only way to obtain some of this information is to conduct a good pilot study.

Pilot studies vary from a brief test of feasibility in a small number of participants to a long trial in hundreds of participants (in preparation for a major multicenter multi-year investment). Pilot studies should be as carefully planned as the main trial, with clear objectives and methods. Many pilot studies are focused primarily on determining the **feasibility**, **time required**, and **cost** of recruiting adequate numbers of eligible participants, and discovering if they are willing to accept randomization and can comply with the intervention. Pilot studies may also be designed to demonstrate that planned **measurements**, data collection **instruments**, and **data management** systems are feasible and efficient. For pilot studies done primarily to test feasibility, a control group is generally not included.

An important goal of many pilot studies is to define the optimal **intervention**—the frequency, intensity, and duration of the intervention that will result in minimal toxicity and maximal effectiveness.

Pilot studies are sometimes used to provide estimates of parameters needed to estimate **sample size**. Sound estimates of the rate of the outcome or mean outcome measure in the placebo group, the effect of the intervention on the main outcome (**effect size**), and the statistical **variability** of this outcome are crucial to planning the sample size. In most cases, it's best to obtain these estimates from published full-scale studies of similar interventions in similar participants. In the absence of such data, using estimates from a pilot study may be helpful, but the sample size for pilot studies is usually so small that the calculated effect size and variance are unstable, with very wide confidence intervals.

Many trials fall short of estimated power not because the effect of the intervention is less than anticipated, but because the rate of dichotomous **outcome events** in the placebo group is much lower than expected. This likely occurs because persons who fit the enrollment criteria for a clinical trial and agree to be randomized are healthier than the general population with the condition of interest. Therefore, it is crucial to determine the rate of the outcome in the placebo group, which may be done by evaluating the placebo group of prior trials with similar participants, or by randomizing participants to placebo in a pilot study.

A pilot study should have a short but **complete protocol** (approved by the institutional review board), data collection forms, and analysis plans. Variables should include the typical baseline measures, predictors, and outcomes included in a full-scale trial, but also estimates of the number of participants available or accessible for recruitment, the number who are contacted or respond using different sources or recruitment techniques, the number and proportion eligible for the trial, those who are eligible but refuse (or say they would refuse) randomization, the time and cost of recruitment and randomization, and estimates of adherence to the intervention and other aspects of the protocol, including study visits. It is usually helpful to "debrief" both participants and staff after the pilot study to obtain their views on how the trial methods could be improved.

A good pilot study requires substantial time and can be costly, but markedly improves the chance of funding for a major clinical trial and the likelihood that the trial will be successfully completed.

CONDUCTING A CLINICAL 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 can be underpowered or biased. Strategies for **maximizing follow-up and adherence** are outlined in Table 11.2.

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 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, giving them pill containers labeled with the day of the week, or sending reminders to their cell phones.

There is also a need to consider how best to **measure adherence** to the **intervention**, using such approaches as self-report, pill counts, pill containers with computer chips that record when the container is opened, and serum or urinary metabolite levels. This information can

TABLE 11.2 MAXIMIZING FOLLOW-UP AND ADHERENCE TO THE PROTOCOL

PRINCIPLE	EXAMPLE
Choose participants who are likely to be adherent to the intervention and protocol	Require completion of two or more visits before randomization
	Exclude those who are non-adherent in a pre-randomization run-in period
	Exclude those who are likely to move or be noncompliant
Make the intervention simple	Use a single tablet once a day if possible
Make study visits convenient and enjoyable	Schedule visits often enough to maintain close contact but not frequently enough to be tiresome
	Schedule visits in the evening or on weekends, or collect information by phone or e-mail
	Have adequate and well-organized staff to prevent waiting
	Provide reimbursement for travel and parking
	Establish good interpersonal relationships with participants
Make study measurements painless, useful, and interesting	Choose noninvasive, informative tests that are otherwise costly or unavailable
	Provide test results of interest to participants and appropriate counseling or referrals
Encourage participants to continue in the trial	Never discontinue follow-up for protocol violations, adverse events, or stopping the intervention
	Send participants birthday and holiday cards
	Send newsletters and e-mail messages
	Emphasize the scientific importance of adherence and follow-up
Find participants who are lost to follow-up	Pursue contacts of participants
	Use a tracking service

identify participants who are not complying, so that approaches to improving adherence can be instituted and the investigator can interpret the findings of the study appropriately.

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 or e-mailing the participant the day before each visit, and by reimbursing travel, parking, and other out-of-pocket costs.

Failure to **follow-up** trial participants and measure the outcome of interest can result in biased results, diminished credibility of the findings, and decreased 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% (13). 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 (14).

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** (see "Analyzing the Results" in this chapter). In many trials, participants who violate the protocol by enrolling in another trial, missing study visits, or discontinuing the study intervention 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 results in more frequent discontinuation of the study medication in those on active treatment compared to those on placebo. If participants who discontinue study medication are not continued in follow-up, this can bias the findings if the side effect is associated with the main outcome or with a serious adverse event (SAE).

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, e-mail address, and telephone number of one or two family members or close acquaintances who will always know where the participant is. In addition to enhancing the investigator's ability to assess vital status, the ability to contact participants by phone or e-mail may give him access to proxy outcome measures from those who refuse to come for a visit at the end. The Heart and Estrogen/Progestin Replacement Study (HERS) trial used all of these strategies: 89% of the women returned for the final clinic visit after an average of 4 years of follow-up, another 8% had a final telephone contact for outcome ascertainment, and information on vital status was determined for every one of the remaining participants by using registered letters, contacts with close relatives, and a tracking service (15).

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. Lengthy and stressful visits can deter some participants from attending. Participants are more likely to return for visits that involve noninvasive tests, such as computed tomography scans, 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 improved by making the trial experience positive and enjoyable for 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 (unless they are specialized research tests that are not yet established for clinical practice); sending newsletters, text messages, or e-mail notes of appreciation; hosting social media sites; sending holiday and birthday cards; giving inexpensive gifts; and developing strong interpersonal relationships with enthusiastic and friendly 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 non-adherent, but not high enough to exclude participants who will turn out to have satisfactory adherence.

A **run-in period** may be useful 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), only those who have complied with the intervention (e.g., taken at least 80% of the assigned placebo) are randomized. Excluding non-adherent participants before randomization in this fashion may increase the power of the study and permit a better estimate of the full effects of intervention. However, a run-in period delays entry into the trial, the proportion of participants excluded is generally small, and participants randomized to the active drug may notice a change in their medication following randomization, contributing to unblinding. It is also not clear that a placebo run-in is more effective in increasing adherence than the requirement that participants complete one or more screening visits before randomization. In the absence of a specific reason to suspect that adherence in the study will be poor, it is probably not necessary to include a run-in period in the trial design.

A variant of the **placebo run-in** design 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 run-in** can select participants who tolerate and respond to the intervention; the absence of adverse effects, or the presence of a desired effect of treatment on a biomarker associated with the outcome, can be used as criteria for randomization. For example, in a placebo-controlled trial testing the effect of nitroglycerin on bone mass, the investigators used a 1-week active run-in period and excluded women who stopped nitroglycerin due to headache (16). This design maximized power by increasing the proportion of the intervention group that tolerated the drug and were likely to be adherent. However, the findings of trials using this strategy may not be generalizable to those excluded.

Using an active run-in may also result in underestimation of the rate of adverse effects. A trial of the effect of carvedilol on mortality in 1,094 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 7 died (17). These people were not randomized in the trial, and these adverse effects of drug treatment were not included as outcomes.

Ascertaining and Adjudicating Outcomes

Data to ascertain that an outcome has occurred can come from many sources: self-report, standardized questionnaires, administrative or clinical records, laboratory or imaging tests, special measurements, and so on. 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:

1. Creating clear criteria for the outcome (e.g., a new, persistent neurologic deficit with corresponding lesion on computed tomography or magnetic resonance imaging scan);
2. Collecting the clinical documents needed to make the assessment (e.g., discharge summaries and radiology reports);
3. Having blinded experts review each potential case and judge whether the criteria for the diagnosis have been met.

The adjudication is often done by two experts working independently, then resolving discordant cases by discussion between the two or by a third expert. However, involving multiple experts in adjudication can be expensive, and for straightforward outcomes in smaller studies it may be sufficiently accurate to have a single investigator carry out the adjudication. The important thing is that anyone involved in collecting the information and adjudicating the cases be blinded to the treatment assignment.

Monitoring Clinical Trials

Investigators must assure that participants are not exposed to a harmful intervention, denied a beneficial intervention, or continued in a trial if the research question is unlikely to be answered. Each of these three considerations must be monitored during the course of a trial to see if the trial should be stopped early.

- **Stopping for harm.** The most pressing reason to monitor clinical trials is to make sure that the intervention does not turn out unexpectedly to be harmful. If **harm** is judged to be clearly present and to outweigh benefits, the trial should be stopped.
- **Stopping for benefit.** If an intervention is more effective than was estimated when the trial was designed, statistically significant **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.
- **Stopping for futility.** If there is a very low probability 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, for example, but after 4 years there is little difference in the rate of outcome events in the intervention and control groups, the "conditional power" (the likelihood of rejecting the null hypothesis in the remaining time, given the results thus far) becomes very small and consideration should be given to stopping the trial. Sometimes trials are stopped early if investigators are unable to recruit or retain enough participants to provide adequate power to answer the research question, or adherence to the intervention is very poor.

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 for either benefit or harm becomes available during a trial, it may be unethical to continue the trial.

Most clinical trials should include an **interim monitoring plan**. Trials funded by the National Institutes of Health (NIH) generally require interim monitoring, even if the intervention is considered safe (such as a behavioral intervention for weight loss). How interim monitoring will occur should be considered in the planning of any clinical trial. In small trials with interventions likely to be safe, the trial investigators might monitor safety or appoint a single independent data and safety monitor. In large trials and trials in which adverse effects of the intervention are unknown or potentially dangerous, interim monitoring is generally performed by a committee, usually known as the Data and Safety Monitoring Board (**DSMB**), consisting of experts in the disease or condition under study, biostatisticians, clinical trialists, ethicists, and sometimes a representative of the patient group being studied. These experts are not involved in the trial, and should have no personal or financial interest in its continuation. DSMB guidelines and procedures should be detailed in writing before the trial begins. Guidance for developing DSMB procedures is provided by the FDA and the NIH. Items to include in these guidelines are outlined in Table 11.3.

Stopping a trial should always be a careful 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. Statistical tests of significance using one of the methods that compensates for multiple looks at the findings (Appendix 11B) provide important but not conclusive information for stopping a trial. Trends over time and 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 11.2).

There are many statistical methods for monitoring the interim results of a trial. Analyzing the results of a trial repeatedly ("multiple peeks") is a form of multiple hypothesis testing and increases the probability of a type I error. For example, if $\alpha = 0.05$ is used for each interim

TABLE 11.3 MONITORING A CLINICAL TRIAL

- Elements to monitor
  - Recruitment
  - Randomization
  - Adherence to intervention and blinding
  - Follow-up completeness
  - Important variables
    - Outcomes
    - Adverse effects
    - Potential co-interventions
- Who will monitor
  - Trial investigator or a single monitor if small trial with minor hazards
  - Independent data and safety monitoring board otherwise
- Methods for interim monitoring
  - Specify statistical approach and frequency of monitoring in advance
  - Importance of judgment and context in addition to statistical stopping rules
- 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

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 from 5% to about 14% (18). To address this problem, statistical methods for interim monitoring generally decrease the α for each interim test so that the overall α is close to 0.05. There are multiple approaches to deciding how to "spend α" (Appendix 11B).

Analyzing the Results: Intention-to-Treat and Per-Protocol

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 many 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 (19).

One important issue that should be considered in the analysis of clinical trial results is the primacy of the intention-to-treat analytic approach to dealing with "**crossovers**," 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 adhered to the assigned intervention. Intention-to-treat analyses may underestimate the full effect of the treatment, but they guard against more important sources of biased results.

An alternative to the intention-to-treat approach is to perform "**per-protocol**" analyses that include only participants who adhered to the protocol. This is defined in various ways, but often includes only participants in both groups who were adherent to the assigned study medication, completed a certain proportion of visits or measurements, and had no other protocol violations. A subset of the per-protocol analysis is an "**as-treated**" analysis in which only participants who were adherent to the assigned intervention are included. These analyses *seem* reasonable because participants can only be affected by an intervention they actually receive. However, participants who adhere to the study treatment and protocol may be different from those who do not in ways that are related to the outcome. In the Postmenopausal Estrogen-Progestin Interventions (PEPI) trial, 875 postmenopausal women were randomly assigned to four different estrogen or estrogen plus progestin regimens and placebo (20). Among women assigned to the unopposed estrogen arm, 30% had discontinued treatment after 3 years because of endometrial hyperplasia, a precursor of endometrial cancer. If these women were eliminated in a per protocol analysis, the association of estrogen therapy and endometrial cancer would 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. Therefore, substantial discontinuation or crossover 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. For example, in the Women's Health Initiative randomized trial of the effect of estrogen plus progestin treatment on breast cancer risk, the hazard ratio was 1.24 ($P = 0.003$) from the intention-to-treat analysis and 1.49 in the as-treated analysis ($P < 0.001$) (21). If the results of intention-to-treat and per protocol analyses differ, the intention-to-treat results generally predominate for estimates of efficacy because they preserve the value of randomization and, unlike per-protocol analyses, can only bias the estimated effect in the conservative direction (favoring the null hypothesis). However, for estimates of harm (e.g., the breast cancer findings), as-treated or per-protocol analyses provide the most conservative estimates, as interventions can only be expected to cause harm in exposed persons.

Results can only be analyzed by intention-to-treat if follow-up measures are completed regardless of whether participants adhere to treatment. Therefore, this should always be the goal.

Subgroup Analyses

Subgroup analyses are defined as comparisons between randomized groups in a subset of the trial cohort. The main reason for doing these analyses is to discover **effect modification** ("**interaction**") in subgroups, for example whether the effect of a treatment is different in men than in women. 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 randomization. For example, a trial of denosumab to prevent fractures found that the drug decreased risk of non-vertebral fracture by 20% among women with low bone density. Preplanned subgroup analyses revealed that the treatment was effective (35% reduction in fracture risk; $P < 0.01$) among women with low bone density at baseline and that treatment was ineffective in women with higher bone density at baseline ($P = 0.02$ for effect modification) (22). It is important to note that the value of randomization is preserved: The fracture rate among women randomized to denosumab is compared with the rate among women randomized to placebo in each subgroup. Subgroup analyses based on post-randomization factors such as adherence to randomized treatment do not preserve the value of randomization and often produce misleading results.

Subgroup analyses can produce misleading results for several reasons. Being smaller than the entire trial population, there may not be sufficient power to find important differences; investigators should avoid claiming that a drug "was ineffective" in a subgroup when the finding

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. For example, if 20 subgroups are examined, differences in one subgroup at $P < 0.05$ would be expected to occur by chance. To address this issue, 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 (23). Claims about different responses in subgroups should be supported by evidence that there is a statistically significant interaction between the effect of treatment and the subgroup characteristic, and a separate study should confirm the effect modification before it is considered established.

SUMMARY

1. There are several variations on the randomized trial design that can substantially increase efficiency under the right circumstances:
   a. The **factorial design** allows two or more independent trials to be carried out for the price of one.
   b. **Cluster randomization** permits efficient studies of naturally occurring groups.
   c. **Non-inferiority or equivalence trials** compare a new intervention to an existing "standard of care."
   d. **Adaptive designs** increase efficiency by allowing design changes based on interim analyses, for example altering the **dose** of study drug, the **number** of participants, and the **duration** of follow-up.
2. There are also other useful clinical trial designs:
   a. **Time series designs** have a single group with outcomes compared within each participant during periods on and off an intervention.
   b. **Crossover designs** combine within and between group designs to enhance control over confounding (if **carryover effects** are not a problem) and **minimize sample size**.
3. Trials for regulatory approval of **new drugs** are classified as:
   a. **Phase I**, small trials to explore dosage and safety
   b. **Phase II**, medium-sized randomized or time series trials of drug effects at several doses
   c. **Phase III**, large randomized trials to demonstrate that benefits outweigh harms as the basis for FDA approval
   d. **Phase IV**, large post-marketing observational studies to confirm benefits and detect rare adverse effects
4. **Pilot studies** are important steps to help determine **acceptability** of interventions and **feasibility**, **size**, **cost**, and **duration** of planned trials.
5. In **conducting a trial**, if a substantial number of study participants **do not adhere** to the study intervention or are **lost to follow-up**, the results of the trial are likely to be underpowered, biased, or uninterpretable.
6. During a trial, **interim monitoring** by an independent **data and safety monitoring board (DSMB)** is needed to assure the **quality** of the study, and to decide if the trial should **stop early** due to evidence of **harm**, **benefit**, or **futility**.
7. **Intention-to-treat** analysis takes advantage of the control of confounding provided by randomization and should be the primary analysis approach for **assessing efficacy**. **Per protocol** analyses, a secondary approach that provides an estimate of the effect size in adherent participants (interpreted with caution), is the most conservative analysis of the harmful effects of treatment.
8. **Subgroup analyses** can detect whether the effect of treatment is modified by other variables; to minimize misinterpretations, the investigator should specify the subgroups in advance, test possible **effect modifications (interactions)** for statistical significance, and report the number of subgroups examined.

APPENDIX 11A
Specifying the Non-Inferiority Margin in a Non-Inferiority Trial

One of the most difficult issues in designing a **non-inferiority trial** is establishing the loss of efficacy of the new treatment that would be unacceptable (7), referred to as "Δ" and often called the **non-inferiority margin**. This decision is based on both statistical and clinical considerations of the potential efficacy and advantages of the new treatment, and requires expert judgment. Here's an example of how this works:

EXAMPLE 11.1 Designing a Study of a New Drug Compared to Warfarin in Patients with Atrial Fibrillation

Warfarin reduces risk for stroke in high-risk patients with atrial fibrillation, so a new drug should be compared to this standard of care. When warfarin is used to reduce the risk of stroke in this situation, it is difficult to dose correctly, requires frequent blood tests to monitor level of anticoagulation, and can cause major bleeding. If a new drug were available that did not have these drawbacks, it could be reasonable to prefer this drug to warfarin, even if its efficacy in reducing risk of stroke was slightly lower.

One approach to setting Δ is to perform a meta-analysis of previous trials of warfarin compared to placebo, and set Δ at some proportion of the distance between the null and lower bound for the treatment effect of warfarin. Alternatively, since studies included in meta-analyses often vary in quality, it may be better to base Δ on the results of the best quality randomized trial of warfarin that has similar entry criteria, warfarin dosage and outcome measures. It is important to set Δ such that there is a high likelihood, taking all benefits and harms into account, that the new therapy is better than placebo (6, 7).

Suppose that a meta-analysis of good-quality trials of warfarin compared to placebo shows that treatment with warfarin reduces the rate of stroke in high-risk patients with atrial fibrillation from 10% per year to about 5% per year (absolute treatment effect = 5%, 95% CI 4–6%). Given the advantages of our new drug, what loss of efficacy is clinically unacceptable? Perhaps an absolute efficacy that is 2% lower than warfarin would be acceptable? In this case, we would declare the new treatment non-inferior to warfarin if the lower limit of the confidence interval around the difference in stroke rates between warfarin and the new treatment is less than 2% (Figure 11.2). In a non-inferiority trial, it is also possible that the new treatment is found to be superior to the established treatment (topmost example in Figure 11.2).

APPENDIX 11B
Interim Monitoring of Trial Outcomes and Early Stopping

Interim monitoring of trial results to decide whether to stop a trial is a form of multiple hypothesis testing, and thereby 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 α is approximately = 0.05. There are multiple statistical methods for decreasing α_i.

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 two disadvantages, however: it requires using an equal threshold for stopping the trial at any interim analysis, and it results in a low α for the final analysis. Most investigators would rather use a more strict threshold for stopping a trial earlier rather than later and use an α close to 0.05 for the final analysis. In addition, this approach is too conservative because it assumes that each test is independent. Interim analyses are not independent, because each successive analysis is based on cumulative data, some of which were included in prior analyses. For these reasons, the Bonferroni method is not generally used.

A commonly used method suggested by O'Brien and Fleming (24) uses a very small α_i for the initial hypothesis test, then gradually increases it for each test such that α_i for the final test is close to the overall α. O'Brien and 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; and N is the total number of tests planned. 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. In addition, this method avoids the awkward situation of getting to the end of a trial and accepting the null hypothesis when the P value is 0.04 or 0.03 but the α_i for the final test is diluted down to 0.01.

A major drawback to the O'Brien–Fleming method 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 become necessary when important trends occur. DeMets and Lan (25) 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, the number of "looks" and the proportion of data to be analyzed at each "look" do not need to be specified before the trial. Of course, for each additional unplanned interim analysis conducted, the final α is a little smaller.

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 probability of rejecting the null hypothesis at the end of the trial, conditioned on the accumulated data. A range of conditional power is typically calculated, first assuming that *Ho* is true (i.e., that any future outcomes in the treated and control groups will be equally distributed) and also assuming that H_a is true (i.e., that outcomes will be distributed unequally in the treatment and control groups as specified by H_a). Other estimates can also be used to provide a full range of reasonable effect sizes. If the

conditional power to reject the null hypothesis across the range of assumptions is low, the null hypothesis is not likely to be rejected and the trial might be stopped.

Examples of two trials that were stopped early are presented in Example 11.2

EXAMPLE 11.2 Two Trials That Have Been Stopped Early

Cardiac Arrhythmia Suppression Trial (CAST) (26). The occurrence of premature ventricular contractions in survivors of myocardial infarction (MI) 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 MI 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 this large and highly statistically significant difference led to the trial being stopped after 18 months.

Physicians' Health Study (27). 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 statistically significant reduction in risk of non-fatal myocardial infarction in the treated group (relative risk = 0.56), but no difference in the number of cardiovascular disease deaths. The rate of cardiovascular disease deaths observed in the study was far lower than expected (88 after 4.8 years of follow-up versus 733 expected), and the trial was stopped because of the beneficial effect of aspirin on risk for nonfatal MI coupled with the very low conditional power to detect a favorable impact on cardiovascular mortality.

REFERENCES

1. Ridker PM, Cook NR, Lee I, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. *N Engl J Med* 2005;352:1293–1304.
2. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials* 1998;19:61–109.
3. Walsh M, Hilton J, Masouredis C, et al. Smokeless tobacco cessation intervention for college athletes: results after 1 year. *Am J Public Health* 1999;89:228–234.
4. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–914.
5. Piaggio G, Elbourne DR, Altman DG, et al. Reporting of non-inferiority and equivalence randomized trials. An extension of the CONSORT Statement. *JAMA* 2006;295:1152–1160.
6. Piaggio G, Elbourne DR, Pocock SJ, et al. Reporting of non-inferiority and equivalence randomized trials. An extension of the CONSORT 2010 statement. *JAMA* 2012;308:2594–2604.
7. Kaul S, Diamond GA. Good enough: a primer on the analysis and interpretation of non-inferiority trials. *Ann Intern Med* 2006;145:62–69.
8. D'Agostino RB Sr., Massaro JM, Sullivan LM, et al. Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statist Med* 2003;22:169–186.
9. Chang M, Chow S, Pong A. Adaptive design in clinical research: issues, opportunities, and recommendations. *J Biopharm Stat* 2006;16:299–309.
10. Chalmers T, Celano P, Sacks H, et al. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358–1361.
11. Pocock S. Current issues in the design and interpretation of clinical trials. *Br Med J* 1985;296:39–42.
12. Nickles CJ, Mitchall GK, Delmar CB, et al. An n-of-1 trial service in clinical practice: testing the effectiveness of stimulants for attention-deficit/hyperactivity disorder. *Pediatrics* 2006;117:2040–2046.
13. Chestnut CH III, Silverman S, Andriano K, et al. A randomized trial of nasal spray salmon calcitonin in postmenopausal women with established osteoporosis: the prevent recurrence of osteoporotic fractures study. *Am J Med* 2000;109:267–276.

14. Cummings SR, Chapurlat R. What PROOF proves about calcitonin and clinical trials. *Am J Med* 2000;109:330–331.
15. 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;280:605–613.
16. Jamal SA, Hamilton CJ, Eastell RJ, Cummings SR. Effect of nitroglycerin ointment on bone density and strength in postmenopausal women. *JAMA* 2011;305:800–805.
17. Pfeffer M, Stevenson L. Beta-adrenergic blockers and survival in heart failure. *N Engl J Med* 1996;334:1396–1397.
18. Armitage P, McPherson C, Rowe B. Repeated significance tests on accumulating data. *J R Stat Soc* 1969;132A:235–244.
19. Friedman LM, Furberg C, DeMets DL. *Fundamentals of clinical trials*, 3rd ed. St. Louis, MO: Mosby Year Book, 1996.
20. 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.
21. Writing group for WHI investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *JAMA* 2001;288:321-333.
22. McClung MR, Boonen S, Torring O, et al. Effect of denosumab treatment on the risk of fractures in subgroup of women with postmenopausal osteoporosis. *J Bone Mineral Res* 2012;27:211–218.
23. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine—Reporting of subgroup analyses in clinical trials. *NEJM* 2007;357:2189–2194.
24. O'Brien P, Fleming T. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549–556.
25. DeMets D, Lan G. The alpha spending function approach to interim data analyses. *Cancer Treat Res* 1995;75:1–27.
26. Cardiac Arrhythmia Suppression Trial (CAST) Investigators. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N Engl J Med* 1989;321:406–412.
27. Physicians' Health Study Investigations. Findings from the aspirin component of the ongoing Physicians' Health Study. *N Engl J Med* 1988;318:262–264.