

# Introduction to Statistical Methods for Clinical Trials

Edited by

Thomas D. Cook

David L. DeMets

Once the hypothesis has been formulated, including what outcome variables will be used to evaluate the effect of the intervention, the next major challenge that must be addressed is the specification of the experimental design. Getting the correct design for the question being posed is critical since no amount of statistical analysis can adjust for an inadequate or inappropriate design. While the typical clinical trial design is usually simpler than many of the classical experimental designs available (Cochran and Cox 1957; Cox 1958; Fisher 1925; Fisher 1935), there are still many choices that are being used (Friedman et al. 1998).

While the concept of a randomized control is relatively new to clinical research, starting in the 1950's with first trials sponsored by the Medical Research Council, there are, of course, examples in history where a control group was not necessary. Examples include studies of the effectiveness of penicillin in treating pneumococcal pneumonia and a vaccine for preventing rabies in dogs. These examples, however, are rare and in clinical practice we must rely on controlled trials to obtain the best evidence of safety and efficacy for new diagnostic tests, drugs, biologics, devices, procedures, and behavioral modifications.

In this chapter, we first describe the early phase trial designs that are used to obtain critical data with which to properly design the ultimate definitive trial. As discussed, once a new intervention has been developed, it may have to go through several stages before the ultimate test for efficacy. For example, for a drug or biologic one of the first challenges is to determine the maximum dose that can be given without unacceptable toxicity. This is evaluated in a *phase I* trial. A next step is to determine if the new intervention modifies a risk factor or symptom as desired, and to further assess safety. This may be accomplished through a *phase II* trial or a series of phase II studies. Once sufficient information is obtained about the new intervention, it must be compared to a control or standard intervention to assess efficacy and safety. This is often referred to as a *phase III* trial. Trials of an approved treatment with long-term follow-up of safety and efficacy are often called *phase IV* trials. (See Fisher (1999) and Fisher and Moyé (1999) for an example of the U.S. Food and Drug Administration (FDA) approval process.) While these phase designations are somewhat arbitrary, they are still useful in thinking about the progression of trials needed to proceed to the final definitive trial. Once the classical phase I and II designs have been described, we discuss choices of control groups, including the randomized control. The rest of the chapter

will focus on randomized control designs, beginning with a discussion of trials designed to show superiority of the new experimental intervention over a standard intervention. Other trials are designed to show that, at worst, the new intervention is not inferior to the standard to within a predetermined margin of indifference. These trials can be extremely difficult to conduct and interpret, and we discuss some of the associated challenges. Finally, we address *adaptive designs* which are intended to allow for design changes in response to intermediate results of the trial.

### 3.1 Early Phase Trials

Medical treatments are subjected to a rigorous evaluation process between the time of their conception, preclinical testing, and confirmation with a definitive phase III trial. Most, in fact, do not complete the process. "Every year the pharmaceutical industry develops thousands of new compounds in the hope that some of them will ultimately prove useful in the treatment of human disease" (Ryan and Soper 2005). In the case of potential cancer treatments, for example, only one in ten succeeds (Goldberg 2006). There are a number of possible development pathways that a particular treatment can take. Generally, new compounds are initially subjected to extensive biochemical and pharmacological analysis, progressing to experiments using animal and *in vitro* models and, increasingly, *in silico*, or computer, models. Once these non-clinical studies are completed, the compound may make the transition from "mouse to man" (Schneiderman 1967; Gart et al. 1986). Investigators will then have to rely completely on prospective studies in humans for preliminary clinical information. On the other hand, a new application may be proposed for a treatment about which much is already known. For example, when aspirin was proposed as an anti-clotting agent for those at risk of heart attack, it had already been in use as a pain reliever for more than eighty years with attendant information on side-effects, overdoses, and other risks.

The first clinical problem (that is, one that is directly related to human treatment) is to determine an acceptable dose and form of administration for an agent. Although "dose" usually refers to the quantity of a drug, it may also apply to other regimens and procedures such as amount or duration of therapeutic radiation. Initial dose-setting studies, often with 20 to 50 subjects, are called phase I trials (Storer 1989). They may be undertaken using healthy voluntary or paid subjects when the expected toxicities are minor or, alternatively, extremely ill patients administered fairly toxic therapies for diseases such as cancer, having failed all standard options. Phase II trials (Thall and Simon 1995) are typically small prospective studies that evaluate a therapy's potential and may be randomized or not.

Although the distinctions between study phases are useful, they are continually evolving and new intermediate or combination designs are introduced, so the lines aren't always clear. Proposed phase I/II trials (Hardwick et al. (2003) and Gooley et al. (1994), for example) investigate both toxicity and short-term

### EARLY PHASE TRIALS

efficacy on the same subjects. Phase II/III designs (Gallo et al. 2006; Storer 1990), also termed *adaptive designs*, begin with an exploratory study of efficacy and, if successful, proceed to a larger confirmatory trial incorporating the same subjects. The FDA has even suggested "phase 0" exploratory studies used for therapy would be given to willing cancer patients and healthy volunteers. Metabolic properties and effects on biomarkers would be assessed and imaging studies performed. Despite these hybridizations, we will associate the term "phase I" with dose-finding trials and "phase II" with efficacy screening trials.

#### 3.1.1 Phase I trials

A phase I trial as defined by Meinert (1996) is "[b]roadly, a trial involving the first applications of a new treatment to human beings and conducted to generate preliminary information on safety." Phase I trials of therapies in cancer and other severe diseases usually have the specific goal of finding the *maximum tolerated dose* (MTD) to use in further testing of efficacy. This dose is quantitatively defined as the largest dose of a drug or other therapeutic agent that induces at most a set proportion, usually 1/3, of subjects with predefined toxicities. Ethical considerations require that the MTD is found by giving the first subjects a very low dose and then sequentially escalating the dose in subsequent patients or subjects until toxicities are observed. The sequential nature of traditional phase I trials requires these toxicities to be short-term, usually occurring within four to eight weeks of initial treatment. See Figure 3.1 for an illustration of the phase I approach in which the goal is to estimate the MTD as the 33rd percentile of the dose-toxicity curve, while minimizing the number of subjects receiving toxic doses.

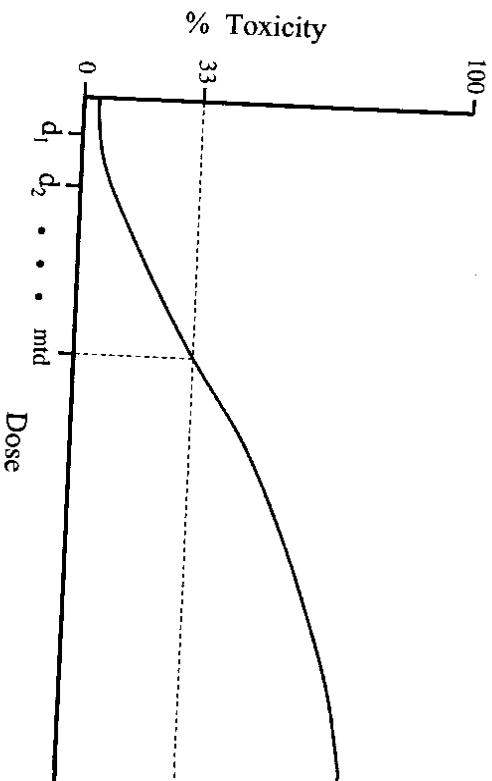


Figure 3.1 Schematic of phase I trial.

Typically, investigators pick a set of three to ten doses for evaluation in a phase I trial based on earlier experience in humans, if prior data exists, or, lacking such data, a fraction of the weight-adjusted dose found to be toxic or fatal to laboratory animals. The intervals between doses are commonly large for low doses and then, for safety's sake, progressively shorten. That is, the second dose might double the first, the third add 50% to the second, and subsequent doses increase by smaller amounts. The details vary between trials but the general pattern of diminishing intervals is known inscrutably as the "Fibonacci method" (it does not necessarily correspond to the mathematical sequence). It is also sensible to specify an even lower level than the starting dose in the trial protocol in the event that the latter unexpectedly proves toxic. In addition to the dose levels, investigators must formulate definitions for adverse events severe enough so that their presence would curtail the use of the treatment. These are called *dose-limiting toxicities* (DLTs).

The so-called *traditional design*, used for cancer drugs and still common despite generally preferable alternatives, uses a simple algorithm applied to cohorts of size three. The first three subjects are assigned to the starting dose; if none experience DLTs, three additional subjects are given the next higher dose. If one subject has a DLT then three more subjects are given the same dose and if none of these has another (so that one of six subjects experiences a DLT at that dose) the next cohort is given the next higher dose; otherwise (that is, if at least two out of three or at least two out of six subjects have DLTs) the trial stops. The algorithm is repeated until either the DLT criterion is met or the maximum dose is reached. The dose below the one at which excessive toxicity is observed is called the MTD.

The traditional design has no fixed sample size—the maximum is six times the number of doses. Its most serious drawback is that because there are multiple opportunities for stopping, on average the MTD can have a DLT rate substantially less than 1/3. Storer (1989) shows that this understimation varies with the unknown dose-response curve and can be substantial. He presents a similar but more appropriate algorithm, the *modified design*, in which, again, cohorts of three subjects are used: if none have a DLT then the next cohort receives a higher dose; if one has a DLT then the current dose is maintained; and if more than one DLT is observed then the dose is *decreased*. The trial proceeds until a preset sample size is achieved. The recommended MTD could be the most recent dose at which excessive toxicity was not observed or, more efficiently, estimated from a logistic regression or perhaps a Bayesian model. Also, cohort sizes of one or two can be used until an DLT is observed in order to facilitate rapid initial dose escalation when this is deemed safe.

There are many variations of this design that can implemented to suit particular circumstances. At the first de-escalation, increments can be halved in order to refine the MTD estimate; information from sub-dose-limiting toxicities can be used not to de-escalate the dose but to slow its increase, for example, to enlarge the cohort size after a sub-DLT is observed; similarly,

toxicity scores can be created that equate a sub-DLT with a fraction of a DLT and exceptionally severe toxicities with more than a unit DLT.

In some circumstances it may be possible to speed up dose escalation by assigning some subjects more than one dose. Simon et al. (1997), in what they called accelerated titration, proposed intra-subject dose escalation. They suggested that if a subject in the first portion of a study does not experience a DLT then he or she could be given a second higher dose. Once an initial estimate of an MTD is made then another group of subjects is assigned to it as their sole dose for confirmation. Obviously, this design is limited to situations where multiple dose assignment to the same subject is feasible.

A fundamentally different approach to dose-escalation can be implemented by using parametric models to dynamically estimate DLT probabilities as a function of dose. This requires prior assumptions regarding likely values of the model parameters and so is essentially a Bayesian model, the earliest and most thoroughly investigated form of which is known as the *continual reassessment method* (CRM) (see O'Quigley et al. (1990), and Garrett (2006) for a tutorial and summary of further developments). Other Bayesian designs such as *Escalation with Overdose Control* (EWOC Babb et al. (1998)) focus on minimizing toxicities to subjects. See Babb and Rogatko (2004) for an overview of Bayesian methods in phase I trials.

The CRM for binary toxicities can be briefly described as follows. Denote the probability of a toxicity at dose  $d_{[j]}$  as  $F(d_{[j]}, \beta)$ , where  $\beta$  is a vector of low (one or two) dimension. (Even though a single scalar parameter is an unrealistic way to characterize a typical dose response curve, it is useful here because of the small sample size and limited goal of estimating the MTD rather than the entire curve.) Let  $p(\beta)$  be a prior distribution for the dose-response parameter, whose choice is discussed below. Then, if  $y_i$  toxicities are observed at dose  $d_{[j]}$ , the likelihood for  $\beta$  after  $n$  dose cohorts is

$$L_n(\beta) = \prod_{i=1}^n F(d_{[j]}, \beta)^{y_i} [1 - F(d_{[j]}, \beta)]^{n_i - y_i}$$

where  $n_i$  is the number of subjects receiving dose  $d_{[j]}$ . The posterior distribution of  $\beta$  is proportional to  $p(\beta)L_n(\beta)$  and its mean or mode can be used to estimate the MTD. The next subject, or subjects, is then simply assigned the MTD. That is, each subject receives what is currently estimated to be the best dose. CRM sample sizes are generally fixed in advance. The final MTD estimate would merely be the dose that would be given the next subject were there to be one. Use of a vague prior distribution for  $\beta$  reduces to maximum likelihood estimation, requiring algorithmic dose modification rules for use in early subjects, where little information is available.

The CRM shares several advantages with many other model-based designs. It unifies the design and analysis process as described above, subjects are assigned to current best estimates. It is very flexible, allowing predictors, sub-dose limiting toxicities, and incomplete information to be incorporated. McGinn et al. (2001) illustrated the latter feature in a trial of radiation dose

escalation in pancreatic cancer. Instead of waiting for full follow-up on a cohort of subjects treated at the current dose, they used the time-to-event CRM (TTTE-CRM) method of Cheung and Chappell (2000) to utilize interim outcomes from subjects with incomplete follow-up to estimate the dose to assign the next subject. Full follow-up is used at the trial's end to generate a final estimate of the MTD. They note that trial length is shortened without sacrificing estimation accuracy, though at the cost of logistical complexity.

Bayesian methods can build on the results of previous human exposure, if any, and in turn provide prior distributions for future studies. A disadvantage of model-based designs compared to algorithmic ones is their lack of transparency, leading clinicians to think of them as "black box" mechanisms, yielding unpredictable dose assignments.

A phase I trial's operating characteristics have ethical and scientific implications, so the prior distribution must be sufficiently diffuse to allow data-driven dose changes but strong enough to disallow overly aggressive escalation. The latter can be restricted by forbidding escalation by more than one dose level per cohort and by conservatively setting the initial dose at less than the prior MTD. Therefore, since prior distributions used for the CRM and similar methods are often chosen based on the operating characteristics of the designs they produce in addition to, or instead of, scientific grounds, one might use a different prior distribution for the analysis.

### 3.1.2 Phase II Trials

A phase II trial is a small study of efficacy intended to provide experience with a treatment and its administration in order to inform the decision to conduct a larger trial and, if this is favorable, to inform its planning. "Small" is a flexible term here, depending on event frequency among other factors, but sample sizes in most phase II trials are sixty or less. The outcome is an indicator of treatment effectiveness. Because most phase II studies are performed in one or two stages, their outcomes usually require short to moderate follow-up, often a year or less, but these are intended as guidelines rather than absolute rules. Phase II studies in neurology may look for improvement in clinical symptoms by, say, six months; in cancer, tumor shrinkage (yes/no) and time to tumor progression (as a failure time variable, with maximum follow-ups of 6 to 24 months) are common. Only in the most severe illnesses is overall survival feasible as a primary outcome.

In their simplest form as small one-arm studies with short term outcomes, phase II trials have a long history. Gehan (1961) was one of the first to describe such trials and to formalize them into two subtypes, preliminary and follow-up. A preliminary trial is intended to screen a treatment for initial evidence of efficacy. Its minimum sample size is sometimes given as fourteen using the following argument. Suppose we have the following pair of hypotheses,

$$\begin{aligned} H_0: & \pi_T = 0.20 \\ H_1: & \pi_T > 0.20, \end{aligned}$$

### EARLY PHASE TRIALS

where  $\pi_T$  is the probability of tumor response or other measure of treatment efficacy, and let  $y$  denote the number of observed successes in our initial sample. One approach is to use a procedure that controls the type II error rate, i.e., the probability that we fail to reject  $H_1$  when it is true. This approach is proper if we want to ensure that we do not prematurely discard promising therapies.

Thus if we wish to use the decision rule that dictates that we pursue further study of a compound unless  $y = 0$ , a type II error is made if  $y = 0$ , yet  $\pi_T > 0.20$ . With a sample of  $n$  subjects, the type II error rate is  $\Pr\{y = 0 | \pi_T\} = (1 - \pi_T)^n$ . We note that for  $n=13$ , this probability is  $0.8^{13} = 0.055$ , and for  $n = 14$ , it is  $0.8^{14} = 0.044$ . Therefore, the design that ensures a type II error rate less than 0.05 requires  $n \geq 14$ . Using the Gehan design, if  $y > 0$ , an additional cohort of subjects is treated and a different decision rule is employed at the second stage.

Trials of this nature in diseases for which no treatment exists (or advanced stages of diseases, at which patients have exhausted available treatments) are sometimes known as phase IIA trials. Trials comparing outcomes to existing treatments having known efficacy can be classified as phase IIB trials, although this distinction is not necessarily sharp.

Often, several phase II trials using the same or similar treatments are carried out at different institutions. This compensates for their small sizes and provides a more varied experience with a particular approach, allowing for a more informed choice of the optimal mode of treatment to be used in a subsequent phase III trial. Simon et al. (1985) point out possible inter-institutional variations in patient selection, definition of or ability to detect outcomes, drug dose and schedule and other aspects of the treatment, and sample size. They therefore advocate randomized phase II trials. They also point out that in such studies it may be practical to investigate several schedules (different doses, timings, and/or modes of administration), and proposed designs with several arms, assuming that one has effectiveness probability equal to  $\pi + \delta$  and the rest have effectiveness probability  $\pi$ . They give a table of sample sizes yielding a 90% chance of selecting the best treatment. For example, for  $\pi = 0.2$ ,  $\delta = 0.15$ , and three schedules, a total sample size of 132 (44 per schedule) gives a 90% chance that the schedule observed to have the best efficacy proportion is truly the best one. One can show that randomization increases the sample size required to detect a given difference between two groups, at the same power and type I error, by a factor of four over that for a trial using a historical control; however, this inflation is in part artifactual because it assumes the historical control rate to be known exactly. Nonetheless, many researchers believe that the elimination of bias using randomization justifies its extra costs: randomized phase II trials are becoming increasingly common in many disciplines.

Fleming (1982) formalized the process of deciding whether to conduct further research on a treatment by proposing three hypotheses for binary mea-

sure of efficacy:

$$\begin{aligned} H_0: & \pi_T \leq \pi_1 \\ H_1: & \pi_1 < \pi_T < \pi_2 \\ H_2: & \pi_T \geq \pi_2, \end{aligned}$$

where for a phase II A study we might take  $\pi_1 = 0.05$ ,  $\pi_2 = 0.2$  (among other possibilities). The sample proportion is compared to two critical values, determined by prespecifying bounds on the type I error probabilities. Two values,  $\alpha_1$  and  $\alpha_2$ , are chosen so that  $P(H_0 \text{ rejected} | H_0 \text{ true}) \leq \alpha_1$  and  $P(H_2 \text{ rejected} | H_2 \text{ true}) \leq \alpha_2$ . Rejection of  $H_2$  is evidence of the treatment's lack of efficacy. Rejecting  $H_0$  indicates promise and consideration of a phase III or another phase II trial. Failure to reject one of these two hypotheses suggests that further investigation may be required.

Preliminary and follow-up stages of phase II studies are increasingly integrated into a single trial with an interim analysis. At a mild cost in complexity, this has the administrative advantage of generating a single protocol needing approval only once by each regulatory committee. It also allows continual subject accrual and obviates the logistical problems of shutting down one study and starting another. Finally, a formal two-stage design allows the joint statistical properties of the entire process to be evaluated. These last will be illustrated by a simple one arm study with a binary outcome and an analysis halfway through.

**Example 3.1.** Researchers at the University of Wisconsin Cancer Center, in a protocol called RO 9471, examined the effects of concurrent hyperthermia using microwaves during radiation treatment of eye tumors. Although the primary outcome was toxicity to eye tissues, it was a single-dose study and, therefore, the desired design features were similar to those of a phase II trial for efficacy and the goal was to show that the probability of toxicity,  $\pi_T$ , is small. The hypotheses were

$$\begin{aligned} H_0: & \pi_T = 0.3 \\ H_1: & \pi_T < 0.3 \end{aligned}$$

with type I error, the probability of erroneously rejecting  $H_0$ , at most 0.05. This can be achieved in a one-stage design with a sample of 29 subjects, rejecting  $H_0$  if there are four or fewer experiencing toxicity. The type I error rate, given  $\pi_T = 0.3$ , is  $0.0260 + 0.0093 + 0.0024 + 0.0004 + 0.0000 = 0.038$  via the binomial formula. By a similar calculation, the power to detect  $H_1$  at  $\pi_T = 0.1$  is 0.84.

Now consider the properties under  $H_0$  of a trial with 28 subjects equally divided into two stages as shown in Figure 3.2. Using this design, if no events are observed among the 14 subjects in the first stage, we can stop and reject  $H_0$  (region A). Similarly, if we observe at least 5 events in stage one we can stop but not reject  $H_0$ . If between one and four events are observed, we continue to stage two (region B). If the total number of events after the second stage is no more than four, we reject  $H_0$  (region C), otherwise we cannot reject  $H_0$  (re-

gion D). Table 3.1 gives the probabilities of rejecting and failing to reject  $H_0$

Region	Action	Number of Events in First Cohort					
		0	1	2	3	4	$\geq 5$
A	Reject $H_0$ , stop at stage 1						
B	Fail to reject $H_0$ , stop at stage 1						
C	Reject $H_0$ after stage 2						
D	Fail to reject $H_0$ after stage 2						

Figure 3.2 Example schematic of a phase II trial.

at the end of each stage, under the null and the alternative hypotheses. The probability of erroneously rejecting  $H_0$  is  $0.043 + 0.007 = 0.05$ . (The calculations are more complex using a failure time outcome in which subjects who are accrued in the first stage can continue to be followed during the second.) Under the alternative hypothesis, there is a 23% probability of stopping and rejecting  $H_0$  after the first stage and a 63% probability of rejecting  $H_0$  after the second stage, yielding 86% power.

Note that the investigators also utilized deterministic curtailment (discussed in Chapter 10), stopping the trial at a point in which its eventual outcome was certain. Regions B and D in Figure 3.2 indicate situations in which, during or after the first stage, they knew that the trial's outcome would be to fail to reject the null hypothesis. There is a 41.7% chance under  $H_0$  of ending the trial after fourteen or fewer subjects. The chances of stopping for the same reason at particular points during the second stage or of stopping near the end of the second stage because five or more events are unobtainable (so rejecting  $H_0$  is certain) are easily calculated.

This example shows a situation in which the power, type I error rate, and maximum sample size for a two-stage design are all about the same as those for

Table 3.1 Probabilities corresponding to the regions in Figure 3.2.

Region	Probability	Probability
	Under $H_0$ ( $\pi = 0.3$ )	Under $H_1$ ( $\pi = 0.1$ )
A	0.007	0.23
B	0.417	0.009
C	0.043	0.63
D	0.533	0.13

the corresponding single stage trial. The former, however, offers the substantial possibility of greatly shortening study duration by stopping early, offering ethical and logistical advantages.  $\square$

Although the two-stage phase II designs discussed here are simple, many refinements are possible such as the use of three or more stages. Failure time outcomes can be used in which case the length of follow-up at each stage will influence total trial duration (Case and Morgan 2001). Thall and Simon (1995) discuss Bayesian phase II designs. As was the case with phase I trials, Bayesian considerations may inform the analysis of a study even when they did not motivate its design.

As a further refinement on two- (and, in principle, three- or more) stage trials Simon (1989) describes designs in which the numbers of subjects in each stage may be unequal. Subject to fixed type I rates (denoted by  $\alpha$ , usually specified to be 5% or 10%) and type II error rates ( $\beta$ , usually between 5% and 20%), many designs satisfy the criteria

$$\Pr(\text{Reject } H_0 | \pi = \pi_1) \leq \alpha$$

and

$$\Pr(\text{Reject } H_0 | \pi = \pi_2) \geq 1 - \beta$$

where  $\pi_1$  is the value of  $\pi$  under  $H_0$  and  $\pi_2$  is a possible value under  $H_1$ . Two strategies for choosing subject allocation between two stages are so-called *optimality* and the *minimax* criterion. Both have been widely applied. Optimality minimizes the expected sample size under the null hypothesis, while minimax refers to a minimization of the maximum sample size in the worst case. Because these goals conflict, optimal designs tend to have large expected sample sizes. Jung et al. (2001) present a graphical method for balancing the two goals. In practice, equal patient allocation to the two stages often serves as a simple compromise. Also, the actual sample size achieved at each stage may deviate slightly from the design. This is particularly true for multi-institutional phase II trials, in which there may be a delay in ascertaining the total accrual and in closing a stage after the accrual goal is met.

### 3.2 Phase III/IV Trials

The typical phase III trial is the first to definitively establish that the new intervention has a positive risk to benefit ratio. Trials intended to provide additional efficacy or safety data after the new intervention has been approved by regulatory agencies are often referred to as phase IV trials. The distinction is not always clear, however, and we shall focus our discussion on phase III trials since most of the design issues are similar for phase IV trials.

#### 3.2.1 Types of Control Groups

As described earlier, there are a number of potential control groups that could be used in a phase III trial. We shall discuss three: the historical control, the concurrent control, and the randomized control.

##### Historical Controls

One of the earliest trial designs is the historical control study, comparing the benefits and safety of a new intervention with the experience of subjects treated earlier using the control. A major motivation for this design is that all new subjects can receive the new intervention. Cancer researchers once used this design to evaluate new chemotherapy strategies (Gehan 1984). This is a natural design for clinicians since they routinely guide their daily practice based on their experience with the treatment of previous patients. If either physicians or patients have a strong belief that the new intervention may be beneficial, they may not want to enroll patients in a trial in which they may be randomized to an inferior treatment. Thus, the historical control trial can thereby alleviate these ethical concerns. In addition, since all eligible patients receive the new intervention, recruitment can be easier and faster, and the required sample size roughly half that of a randomized control trial, thereby reducing costs as well. One of the key assumptions, however, is that the patient population and the standard of care remain constant during the period in which such comparisons are being made.

Despite these benefits, there are also many challenges. First, historical control trials are vulnerable to bias. There are many cases in which interventions appeared to be effective based on historical controls but later were shown not to be (Moertel 1984). Byar describes the Veterans Administration Urological Research Group trial of prostrate cancer, comparing survival of estrogen treated patients with placebo treated patients.<sup>1</sup> During this trial, there was a shift in the patient population so that earlier patients were at higher risk. Estrogen appeared to be effective in the earlier high risk patients but not so in the patients recruited later who were at lower risk. A historical control study would have been misleading due to a shift in patient referral patterns. Pocock (1977b) describes a series of 19 studies that were conducted immediately after

<sup>1</sup> Veterans Administration Cooperative Urological Research Group (1967), Byar et al. (1976)

an earlier study of the same patient population and with the same intervention. In comparing the effects of the intervention from the consecutive trials in these 19 cases, 4 of the 19 comparisons were "nominally" significant, suggesting that the treatment in one trial was more effective in one study than the other, even though the patients were similar and the interventions were identical.

A recent trial of chronic heart failure also demonstrates the bias in historical comparisons (Packer et al. 1996; Carson et al. 2000). The initial PRAISE trial evaluated a drug, amiodipine, to reduce mortality and morbidity in a chronic heart failure population. The first trial, referred to as PRAISE I, was stratified by etiology of the heart failure: ischemic and non-ischemic. Earlier research suggested that amiodipine should be more effective in the ischemic population or subgroup. The overall survival comparison between amiodipine and placebo treated patients, using the log-rank test, resulted in a  $p$ -value of 0.07, almost statistically significant. There was a significant interaction between the ischemic and non-ischemic subgroups, however, with a hazard ratio of 1.0 for the ischemic subgroup and 0.5 for the non-ischemic subgroup. While the result in the non-ischemic group alone might have led to its use in clinical practice, the fact that the result was opposite to that expected persuaded the investigators to repeat the trial in the non-ischemic subgroup alone. In the second trial, referred to as PRAISE II, the comparison of the amiodipine and placebo treated arms resulted in a hazard ratio of essentially 1.0. Of interest here, as shown in Figure 3.3, is that the placebo treated patients in the second trial were significantly superior in survival to the placebo treated patients in PRAISE I. No adjustment for baseline characteristics could explain this phenomenon.

In addition, to conduct an historical control trial, historical data must be available. There are usually two main sources of historical data; a literature resource or a data bank resource. The literature resource may be subject to publication bias where only selected trials, usually those with a positive significant benefit, are published and this bias may be introduced into the historical comparison. In addition, to conduct a rigorous analysis of the new intervention, access to the raw data would be necessary and this may not be available from the literature resources. Thus, researchers often turn to existing data banks to retrieve data from earlier studies. Quality of data collection may change over time as well as the definitions used for inclusion criteria and outcome variables.

Even if data of sufficient quality from earlier studies are available, caution is still required. Background therapy may have changed over time and diagnostic criteria may also have changed. For example, international classification of disease changes periodically and thus there can be increases or decreases in disease prevalence. For example, the seventh, eighth, and ninth revisions of the international classification system resulted in a 15% shift in deaths due to ischemic heart disease. Besides changes in diagnostic criteria and classification codes, disease prevalence can also change over time as shown in Figure 3.4.

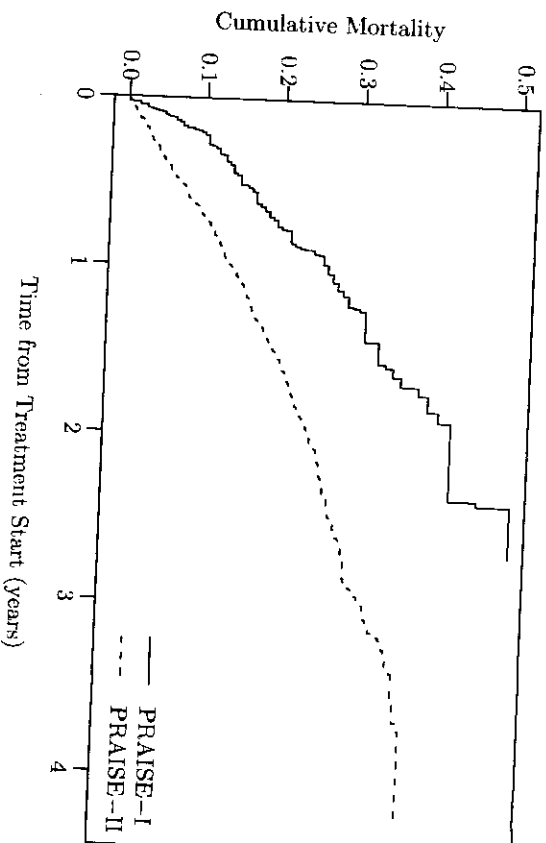


Figure 3.3 PRAISE I vs. PRAISE II: all-cause mortality for placebo groups by study

In this case, historical comparisons of two periods for cardiovascular disease might indicate that more recent interventions were successful. The difference in rates may, however, be due solely to a change in prevalence.

Historical comparisons can be useful to a limited extent but may have bias. Thus, such comparisons may be used early in the assessment of a new intervention to suggest further research and may be used to help design definitive trials.

#### Concurrent Controls

The concurrent control trial compares the effect of the new intervention with the effect of an alternative intervention applied at some other site or clinic. This design has many of the advantages of the historical control trial but eliminates some of the sources of bias. The primary advantage is that the investigator can apply the new intervention to all participants and only half as many new participants are needed, compared to the randomized control trial. Thus, recruitment is easier and faster. The biases that affect historical controls due to changes in definitions or diagnostic criteria, background therapy, and changing time trends in disease prevalence are minimized if not eliminated. These types of comparisons are somewhat common in medical care as success rates of various institutions in treating patients are often compared and evaluated.

The key problem with the concurrent control trial is selection bias, both from patients or participants and the health care provider. Referral patterns

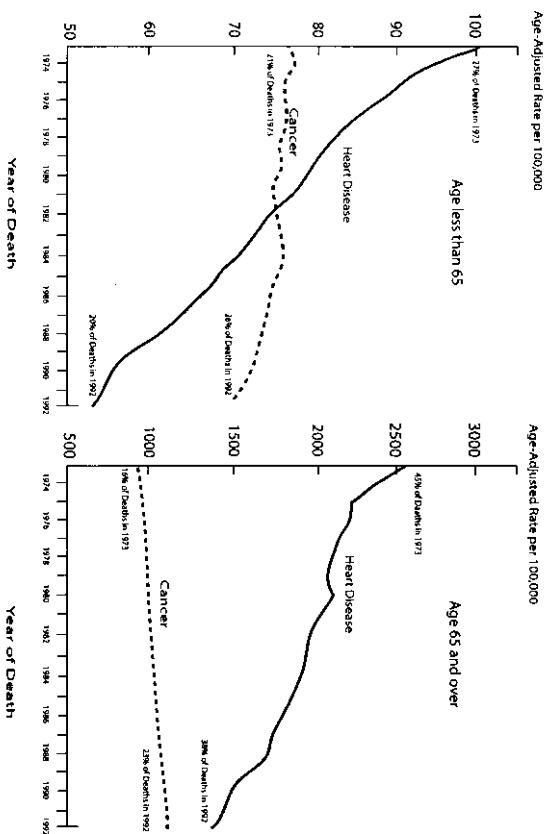


Figure 3.4 *Cancer and heart disease deaths. Cancer and heart disease are the leading causes of death in the United States. For people less than 65, heart disease death rates declined greatly from 1973 to 1992, while cancer death rates declined slightly. For people age 65 and older, heart disease remains the leading killer despite a reduction in deaths from this disease. Because cancer is a disease of aging, longer life expectancies and fewer deaths from competing causes, such as heart disease, are contributing to the increase in cancer incidence and mortality for those age 65 and older. Reprinted with permission from Mchintosh (1995).*

are not random but based on many factors. These may include whether the institution is a primary care facility or a tertiary referral center. Patient mix would be quite different in those two settings. Patients may choose to get their care from an institution because of its reputation or accessibility. With the development of large multidisciplinary health care systems, this source of bias may not be as great as it otherwise might be. Even within such systems, however, different clinics may have different expertise or interest and select patients accordingly.

The key to any evaluation of two interventions is that the populations are comparable at the start of the study. For a concurrent control trial, one would have to examine the profile of risk factors and demographic factors. There are many other important factors, however, that may not be available. Even with a large amount of baseline data, establishing comparability is a challenging task. In the previous section, the PRAISE I & II example indicated it was not possible to explain the difference in survival between the same placebo subgroups in the two back to back studies by examining baseline risk factors. Thus, covariate adjustment cannot be guaranteed to produce valid treatment comparisons in trials using concurrent controls.

### Randomized Control Trials

As indicated earlier, the randomized control trial is viewed as the "gold standard" for evaluating new interventions. The reason for this, as summarized in Table 3.2, is that many of the sources of bias present in both the historical and concurrent control trials are minimized or eliminated (Friedman et al. 1985).

Table 3.2 *Sources of bias as a function of the control group.*

Design	Sources of Bias
Randomized	Chance
Concurrent (Non-randomized)	Chance & Selection Bias
Historical (Non-randomized)	Chance, Selection Bias, & Time Bias

In a randomized control trial, all intervention arms are assessed at the same time under the same conditions. Since a randomized control trial assigns participants to either the standard control or the new intervention at random, selection bias is eliminated. Neither participant nor health care provider can influence which of the two or more interventions are received. This influence may be conscious or subconscious but in either case, the selection bias can be substantial. A randomized control trial minimizes this possibility. In addition, the process of randomization tends to produce, on average, comparable groups of patients in each of the intervention arms. This is true both for measured and unmeasured risk factors. Investigators typically present a baseline covariate table in their trial results publication, demonstrating comparability for those risk factors that were measured.

Another benefit of randomization is that the process of randomization justifies the common statistical tests used to evaluate the interventions (Byar et al. 1976; Kempthorne 1977). Through randomization, the validity of the statistical tests can be made without invoking assumptions about the distribution of baseline variables. Often these assumptions are not strictly true or, at least, are difficult to establish.

Table 3.3 provides an example of the degree of bias that can occur with different designs (Chalmers et al. 1977; Peto et al. 1976). A series of randomized and non-randomized trials for the evaluation of anticoagulant therapy in patients with a myocardial infarction (heart attack) were summarized, noting the estimate of the intervention effect in each class of designs. In this series, there were 18 historical control trials, 8 non-randomized concurrent trials, and 6 randomized control trials. Each class of designs involves several hundred or more participants. As shown, the historical control series and the concurrent control series estimate a 50% intervention effect while the randomized control trials give a smaller estimate of 20%. This example demonstrates the consid-

erable bias that non-randomized designs can produce, bias likely due to time trends and patient selection.

Table 3.3 Possible bias in the estimation of treatment effects for published trials involving anticoagulation for patients with myocardial infarction as a function of the control group. The randomized control is the most reliable.

Design	Patients	P<0.05	Observed Effect
18 Historical	900	15/18	50%
8 Concurrent	3000	5/8	50%
6 Randomized	3000	1/6	20%

Despite the advantages of randomization for minimizing bias, some investigators or participants may object to the randomization process. They may believe that half of the participants are being deprived of access to the new intervention that may be beneficial, regardless of the strength of the evidence or lack thereof. There is an ethical imperative for the physician to do what they believe is best for their patient and what might be considered ethical behavior by one investigator might not be for another. Byar et al. (1993) argue that the most ethical behavior, when evidence for the safety and benefit of a new intervention is not well established, is to find out as quickly as possible with the best, most unbiased methods available. The randomized control trial satisfies this requirement according to Chalmers et al. (1983) who, among others, pioneered the use of this design. Of course, if a physician or a participant strongly believes that the new intervention is better or less toxic than the standard being used for comparison, they should not participate in the trial.

### 3.2.2 Common Randomized Control Designs

The designs for randomized control trials that we will describe are, in general, basic and straightforward. More complicated designs are available but have not seen widespread use. Basic designs are also simpler to analyze, requiring fewer assumptions. Other authors have also summarized these designs (Friedman et al. 1985).

#### Parallel Group Design

The most common design used in clinical research is the randomized *parallel group* (or *parallel control*) design depicted in Figure 3.5. Using this design, participants are screened for eligibility, provide informed consent, have baseline information collected, and are then randomized to one of the intervention arms. After a period of follow-up, participants are evaluated for compliance to the interventions, side effects and toxicities, and the primary and secondary

outcome variables. The intervention arms are compared with respect to the outcome variables specified in the protocol.

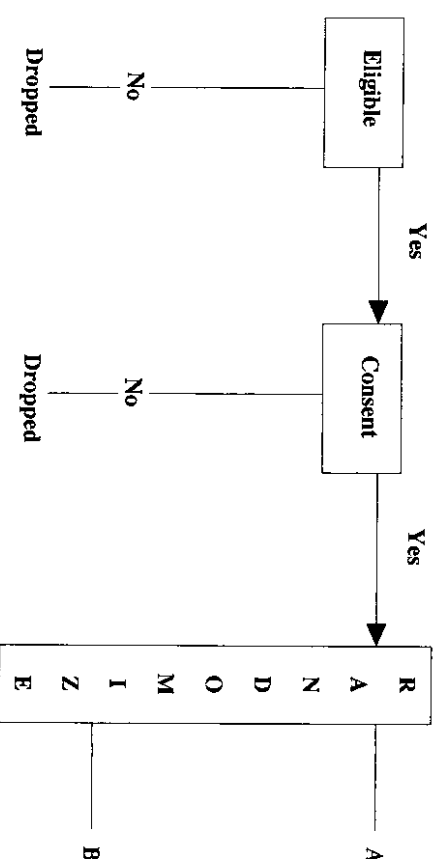


Figure 3.5 Parallel group design.

There are many clinical trials in various fields that have successfully used this trial design.<sup>2</sup> The parallel control design has the advantage of simplicity and can give valid answers to one or two primary questions. For example, the Coronary Drug Project (CDP), one of the early randomized control multicenter clinical trials, compared several intervention strategies with a placebo control arm using a parallel design in a population of men who had recently suffered a heart attack. The primary outcome was mortality with cardiovascular mortality as a secondary outcome. The intervention strategies used various drugs that were known to lower serum cholesterol levels but it was not known whether any could safely lower mortality. The Diabetic Retinopathy Study (DRS) was a trial to evaluate a new laser treatment in a diabetic population to reduce the progression of retinopathy, an eye disease that reduces visual acuity. The primary outcome was visual acuity and a retinopathy score which measures disease progression and is based on photographs of the eye fundus (i.e., back of the eye). The Beta-blocker Heart Attack Trial (BHAT) was a randomized double-blind parallel design trial in a group of individuals having just survived a heart attack, comparing a beta-blocker drug with a placebo control. Again, mortality was the primary outcome variable. The Breast Cancer Prevention Trial (P-1) was a cancer prevention trial evaluating the drug tamoxifen to prevent the occurrence of breast cancer in a population at risk (Fisher et al. 1998). Here, the primary outcome was disease free sur-

<sup>2</sup> The International Steering Committee on Behalf of the MERIT-HF Study Group (1997), Domanski et al. (2002), HDLFP Cooperative Group (1982), The Coronary Drug Project Research Group (1975), The DCCCT Research Group: Diabetes Control and Complications Trial (DCCCT) (1986), Diabetic Retinopathy Study Research Group (1976)

vival, that is, alive without the occurrence of breast cancer. Most phase III trials use this basic design because of its simplicity and utility.

A variation of the parallel design utilizes a *run-in* period. A schema for this design is shown in Figure 3.6. The primary departure from the basic parallel design is that participants are screened into a prerandomization phase to evaluate their ability to adhere to the protocol or to one of the interventions under evaluation. If they cannot adhere to the intervention schedule such as taking a high percentage of the medication prescribed, then their lack of compliance will affect the sensitivity or power of the trial. If a potential participant cannot comply with some of the required procedures or evaluations, then that individual would not be a good candidate for the main trial.

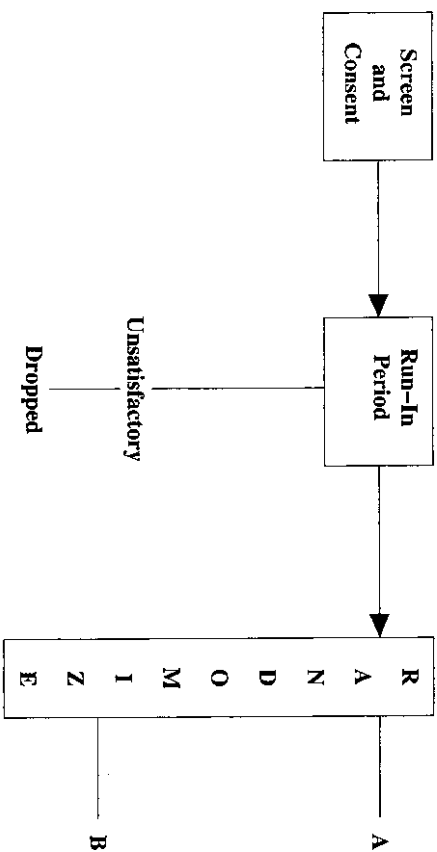


Figure 3.6 Run-in design.

There are many trials that have used a “run-in” period. For example, the Cardiac Arrhythmia Suppression Trial (CAST) used a run-in phase to determine if a patient’s cardiac arrhythmia could be suppressed using one of the three drugs being tested.<sup>3</sup> CAST was a trial involving patients with a serious cardiac arrhythmia, or irregular heartbeat. Individuals with these irregular heartbeats are known to be a high risk for death, usually suddenly and without warning. Researchers developed a class of drugs that would suppress or control these irregular heartbeats, on the theory that this would reduce the risk of sudden death. Not all patients could tolerate these drugs, however, and in some patients the treatment failed to control the arrhythmia. Therefore, to improve the efficiency and power of the main CAST trial, patients were screened to determine both their ability to tolerate these drugs and the susceptibility of the arrhythmia to pharmacological control. If they met the screening criteria, they were randomized into the main trial, either to one of the 3 drugs or to a matching placebo control. Ironically, these drugs were

<sup>3</sup> The Cardiac Arrhythmia Suppression Trial (CAST) Investigators (1989)

## PHASE III/IV TRIALS

shown to be harmful in a patient population who had passed the screening run-in phase with a successful suppression of their arrhythmias.

Another trial, the Nocturnal Oxygen Therapy Trial (NOTT), evaluated the benefit of giving 24 hours of continuous oxygen supplementation, relative to giving 12 hours of nocturnal use, in patients suffering from advance chronic obstructive pulmonary disease (COPD). Potential participants were entered into a run-in phase to establish that their disease was sufficiently stable to ensure that the outcome variables, pulmonary function, could better reflect the treatment effect. A similar strategy was used in the Intermittent Positive Pressure Breathing (IPPB) Trial.<sup>4</sup>

### Randomized Withdrawal Design

A variation in the parallel design is the randomized withdrawal study. Many treatments are shown to be beneficial for a specified period of time, but the required duration of treatment is often not well known. Longer exposure to a drug, for example, may increase the risk of toxicity. If there are no long term benefits, then this risk may not be justified. A randomized withdrawal trial allows design researchers to evaluate the optimal duration of treatment. All participants are given the intervention for a prespecified period of time during which the intervention has been shown to be beneficial. After a predetermined period of exposure, participants are randomly withdrawn from the intervention, as shown in Figure 3.7. The analysis will focus on outcomes starting from the point of randomization. The randomized withdrawal design should be used

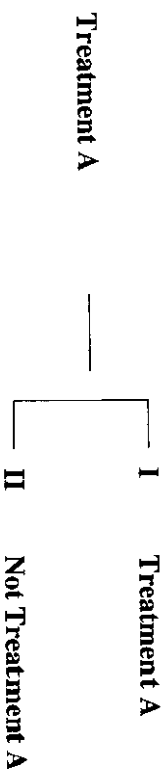


Figure 3.7 Withdrawal design.

more often to give better evidence that can help maximize benefit and minimize long term risk. There are few examples of the randomized withdrawal design.

### Crossover Design

The crossover design has often been used in the early stages of the development of new therapies to evaluate their effects compared to a standard control. Crossover designs are commonly used for studies of analgesic and psychotropic drugs. One of the unique features of this design is that each patient receives both treatments and, hence, serves as his or her own control. In the simple two-period crossover design, as depicted in Figure 3.8, the participants are

<sup>4</sup> The Intermittent Positive Pressure Breathing Trial Group (1983)

randomly divided into two groups and each participant is exposed to the new treatment (A) and to the control (B), each for a prespecified period of time. In Figure 3.8 these time periods are labeled as period I and period II. Group I is exposed to treatment A in period I and treatment B in period II, while group 2 is exposed to treatment B in period I and treatment A in period II.

$H_0$ : A vs. B Scheme

Group	Period I	Period II
1	TRT A	TRT B
2	TRT B	TRT A

Figure 3.8 Two-period crossover design.

The model, according to Brown (1980) and Grizzle (1965), can be written as

$$Y_{ijk} = \mu + \pi_k + \phi_u + \lambda_v + \xi_{ij} + \varepsilon_{ijk},$$

where  $i = 1, 2, j = 1, \dots, n_i$  ( $n_i$  is the number of subjects in group  $i$ ),  $k = \text{I, II}$ ,  $u = \text{A, B}$ , and  $v = \text{A, B}$ . The terms in the model are defined as follows:

- $Y_{ijk}$  = the measurement for subject  $j$  in group  $i$  during period  $k$ ,
- $\mu$  = the overall mean,
- $\pi_k$  = the effect of period  $k$ ,
- $\phi_u$  = the effect of treatment  $u$ ,
- $\lambda_v$  = the carryover effect of treatment  $v$  from period I on the response in period II,
- $\xi_{ij}$  = the subject effect,
- $\varepsilon_{ijk}$  = random error.

In addition,  $E[\xi_{ij}] = E[\varepsilon_{ijk}] = 0$ ,  $\text{Var}(\xi_{ij}) = \sigma_\xi^2$ ,  $\text{Var}(\varepsilon_{ijk}) = \sigma_\varepsilon^2$ , and the  $\xi_{ij}$  and  $\varepsilon_{ijk}$  are assumed to be mutually independent. The carryover effect,  $\lambda_v$ , has a value of 0 for all measurements in period I. Its value in period II is of particular interest and may not be 0. For example, if treatment A cures the disease in period I, then there is no possibility of a response to treatment B in period II. The validity and efficiency of the crossover design depends strongly on whether or not there is any carryover effect. The presentation here follows Brown (1980).

First, we consider at the case in which we assume that  $\lambda_v = 0$ .<sup>5</sup> In this case, we impose the additional constraint that  $\phi_A + \phi_B = 0$ , and the difference between the two treatment effects,  $\delta = \phi_B - \phi_A$ , can be estimated by

$$\hat{\delta}_{co} = \frac{1}{2} [(\bar{Y}_{1,2} - \bar{Y}_{1,1}) + (\bar{Y}_{2,1} - \bar{Y}_{2,2})],$$

<sup>5</sup> Actually we need only assume that  $\lambda_A = \lambda_B$ ; however, if carryover effects are present, it is unlikely that they will be the same for both treatments.

where  $\bar{Y}_{i,k}$  is the average measurement of all subjects in group  $i$  during period  $k$ . It is easy to check that

$$E[\hat{\delta}_{co}] = \delta \quad \text{and} \quad \text{Var}(\hat{\delta}_{co}) = \frac{\sigma_\varepsilon^2}{2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right),$$

where  $\sigma_\varepsilon^2$  can be estimated by the subject differences between periods with  $n_1 + n_2 - 2$  degrees of freedom. As pointed out by Chassan (1970), the efficiency of the crossover design can be compared to that of the randomized parallel control design using simple calculations. Let the number of subjects in each group in the crossover design be  $n$  and the number of subjects in each arm in the parallel design be  $m$ . The variance of the estimator in the crossover trial becomes  $\text{Var}(\hat{\delta}_{co}) = \sigma_\varepsilon^2/n$  and the variance of the estimator in the parallel design experiment will be

$$\text{Var}(\hat{\delta}_p) = \frac{2}{m} (\sigma_\xi^2 + \sigma_\varepsilon^2) = \frac{2\sigma_\varepsilon^2}{m(1-\rho)}$$

where  $\rho$  is the correlation between measurements in period I and period II for a randomly selected subject. The ratio of the variances for the two experiments will be

$$\frac{\text{Var}(\hat{\delta}_{co})}{\text{Var}(\hat{\delta}_p)} = \frac{m}{2n} (1 - \rho).$$

Thus, to achieve the same precision, we require that  $m = 2n/(1 - \rho)$  which depending on  $\rho$ , is at least twice the sample size, and likely much more. Another observation made by Chassan (1970) is that if the analysis of the parallel design experiment was based on change from baseline, then the appropriate estimator, say  $\hat{\delta}_{p(b)}$ , would have variance

$$\text{Var}(\hat{\delta}_{p(b)}) = \frac{4}{m} \sigma_\varepsilon^2.$$

This is four times the variance of the estimator for the crossover experiment when the two have the same sample sizes. Similarly if the estimate is based on equation (2.20) from Chapter 2, the variance is  $2(1 + \rho)\sigma_\varepsilon^2/m$ , which is between two and four times greater than  $\text{Var}(\hat{\delta}_p)$ .

The efficiency results are valid only when  $\lambda_v = 0$ , however. When this assumption is not reasonable, adjustments need to be made that ultimately defeat the purpose of the crossover design. Grizzle (1965) noticed that the hypothesis of no carryover effect can be tested in the simple two-period crossover design. Letting  $\gamma = \lambda_B - \lambda_A$  (i.e., the difference in carryover effect between the two treatments), an unbiased estimator of  $\gamma$  is

$$\hat{\gamma} = (\bar{Y}_{2,1} + \bar{Y}_{2,2}) - (\bar{Y}_{1,1} + \bar{Y}_{1,2}),$$

with variance

$$\text{Var}(\hat{\gamma}) = (4\sigma_\xi^2 + 2\sigma_\varepsilon^2) \left( \frac{1}{n_1} + \frac{1}{n_2} \right).$$

We can estimate  $4\sigma_\xi^2 + 2\sigma_\varepsilon^2$  by estimating the variance of the sum of the two

observations for each individual with  $n_1 + n_2 - 2$  degrees of freedom. Once these estimates are available, we can test the hypothesis of no carryover effect. If it is determined that carryover effect is present, then  $E[\hat{\delta}_{\alpha\beta}] \neq \delta$ . A valid (unbiased) estimate of the treatment effect is  $\bar{Y}_{1,2} - \bar{Y}_{1,1}$ —the estimate from a parallel group trial and which does not make use of the period II data, subverting the entire benefit of the crossover design.

When in doubt about the carryover effect, Grizzle (1965) recommends testing  $\gamma = 0$  at a significance level of 0.1. If this hypothesis is not rejected,  $\hat{\delta}_{\alpha\beta}$  can be used. If it is rejected, only information from the first period should be used. On the other hand, in order to have sufficient power for both this test and the test for treatment effect, one would need a larger sample size and a simple parallel design would be a more efficient choice. Thus, when there is a very strong belief that  $\gamma = 0$ , a crossover design may be used. If there is a possibility of carryover effect, however, one should avoid the crossover design.

These results are a summary of those presented by Brown (1980). Another discussion, arriving at many of the same conclusions, is given by Hills and Armitage (1979). Many other possible designs for crossover trials have been explored. These involve a larger number of treatments or periods and may involve various patterns of treatment assignment. Some of these have been discussed in Koch et al. (1989) and Carriere (1994).

#### Factorial Designs

The factorial design is the most complex of the designs typically used in clinical trials. Using this design, two or more classes of interventions are evaluated in the same study compared to the appropriate standard or control for each. As shown in Figure 3.9 for a two-by-two factorial design, evaluating two interventions A and B compared to a control, there are four cells, each reflecting an intervention strategy. Approximately 25% of the participants are in each cell; that is, 25% receive both interventions AB, 25% receive A and control, 25% receive B and control, and 25% receive both controls. The factorial design allows researchers to evaluate more than one intervention in the same participant population, thus reducing costs and increasing efficiency. Furthermore, as in the Physicians Health Study (PHS) example described later in this section, each treatment comparison may be associated with a different outcome.

	Control	Trt B	Tot
Control	N/4	N/4	N/2
Trt A	N/4	N/4	N/2
Tot	N/2	N/2	N

Figure 3.9 *Balanced 2 × 2 factorial design.*

There are several possible effects of interest including,

1. the overall effect of treatment A on its primary outcome,

#### PHASE III/IV TRIALS

2. the overall effect of treatment B on its primary outcome,
3. the interaction effect, or equivalently, the modification of the effect of treatment A by treatment B or vice versa,
4. the effect of treatment A on its primary outcome, in the presence of treatment B, or
5. the effect of treatment B on its primary outcome, in the presence of treatment A.

In trials using factorial designs solely for efficiency (PHS, for example), interest is likely to be in (1) and (2), the overall effects of each treatment, ignoring the other. Other studies may be conducted to determine if the effect of the two treatments together is different than the effects of the two separately. In this case interest will be in (3), and probably one or both of (4) and (5). If interest were solely in (4) or (5), there would be no reason to use the factorial design.

The typical model used in the analysis of the two-by-two factorial design is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad (3.1)$$

where  $i = 1, \dots, I$ ,  $j = 1, \dots, J$ , and  $k = 1, \dots, K$ .  $Y_{ijk}$  is the outcome of interest for the  $k$ th subject receiving level  $i$  of intervention A and level  $j$  of intervention B. In the design depicted in Figure 3.9,  $I = J = 2$  and  $K = N/4$ . The errors  $\epsilon_{ijk}$  are assumed to be independent and normally distributed with mean zero and variance  $\sigma^2$ .  $\mu$  is interpreted as the overall mean,  $\alpha_i$  is the (additional) effect due to level  $i$  of intervention A,  $\beta_j$  is the (additional) effect due to level  $j$  of intervention B, and  $(\alpha\beta)_{ij}$  is the effect due to the combination of level  $i$  of intervention A with level  $j$  of intervention B. The constraints assumed here are  $\alpha_1 + \alpha_2 = 0$ ,  $\beta_1 + \beta_2 = 0$ , and  $\sum_{i,j} (\alpha\beta)_{ij} = 0$ . If each treatment comparison involved a different outcome, there would be two such models. Note that unless one can *a priori* rule out an effect of, say, treatment B on the outcome associated with treatment A, each model should involve effects for both treatments.

Now consider the marginal effect of treatment A. One school of thought suggests that if the interaction,  $(\alpha\beta)_{ij}$ , is non-zero, then treatment B is an effect modifier for treatment A and analyses of the effect of treatment A must be performed separately for each level of treatment B. This would be technically true if the primary question solely involved point estimation of treatment effects. On the other hand, recall from Chapter 2 that the primary goal of randomized trials is to test hypotheses regarding the effects of the treatments. From this point of view, the tests of the overall effect of treatment A are still valid, regardless of the presence of the interaction. Furthermore, unless the interaction is quite large, so that, for example, there is little or no effect of treatment A in the presence of treatment B, the test for the overall effect of treatment A should still have good power. In fact, there is, in principle, no reason why treatment B should be considered any differently than baseline variables such as age, comorbidities, or concomitant medications, all of which are potential effect modifiers.

In addition, unless the interaction is such that the effect of treatment A is reversed by treatment B, the interaction is likely to be *removable*, in the sense that a transformation can be applied that will make the interaction zero. For example, if the outcome is dichotomous (success or failure), there may be an interaction using the probability scale (the difference in failure rates is a function of the level of treatment B), but the interaction disappears if the log-odds scale is used (example 4.2 in Section 4.6). Thus, the choice of scale may be more important when factorial designs are used than in other trials.

Under the assumption of no interaction, the test of  $H_0: \alpha_1 = 0$  is based on the statistic  $D = (\bar{Y}_{21} - \bar{Y}_{11} + \bar{Y}_{22} - \bar{Y}_{12})/2$ , where  $\bar{Y}_{ij}$  is the mean of the observations in cell  $i, j$ . For robustness against the actual presence of interaction, the best estimate of  $\text{Var}(D)$  is  $16\hat{\sigma}^2/N$ , where  $\hat{\sigma}^2$  is the pooled estimate of  $\sigma^2$  derived from the within-cell variances of the  $Y_{ijk}$  (hence with  $N - 4$  degrees of freedom) rather than the residual mean square error from the additive model. Alternatively, if outcomes are not assumed to be normal, other analyses, stratified by levels of treatment B, can be performed as given in Section 2.3.1. For example, if  $Y_{ijk}$  is dichotomous, a Cochran-Mantel-Haenszel test can be used, or if  $Y_{ijk}$  is ordinal, the Van Elteren test may be appropriate.

If interactions are of interest, for normal data, the  $t$ -test or  $F$ -test corresponding to the interaction term in (3.1) can be performed. Alternatively, for dichotomous data, the Breslow and Day test for interaction (Breslow and Day 1980) may be used, or a logistic regression model may be used, although both of these tests consider only interactions on the log-odds scale. Interaction tests for other types of responses such as failure time outcomes are also available. Note, however, because interactions are intrinsically linked to a particular parameterization, they are necessarily model-based and nonparametric tests are not available. Also note that failure to reject the null hypothesis that there is no interaction does not imply that none exists. In fact, unless the trial is sufficiently large, interaction tests are not likely to be adequately powered (see Section 4.6). Since the sample size required to detect all but the most extreme interactions is many-fold larger than that required for the main effects, doing so will likely more than offset the efficiency gained by using a factorial design.

Next, after the tests of hypotheses have been completed, point estimates will generally be required. Unless the interactions are quite large, in most cases one should report both overall differences by treatment (either adjusted or unadjusted for the potential effect of the other factor)—although the primary difference may be solely in the estimates of standard errors). If desired, the interactions and within-stratum effects may also be reported.

Finally, with multiple factors come multiple tests and multiple testing issues. Adjustments for multiple comparisons should be built into the design of the trial (see Section 11.5). Maintaining power while accounting for multiple tests typically requires increased sample sizes, but in the case of factorial designs, not enough to offset the gain in efficiency.

An example of a factorial design is the Physicians' Health Study (PHS),<sup>6</sup> a

trial evaluating two potential prevention agents, aspirin and beta-carotene. The PHS was a randomized double-blind placebo controlled trial in a population of U.S. male physicians. These participants were randomized to either aspirin, beta-carotene, both, or neither, as shown in Figure 3.9. The primary questions involved only the main effects and each treatment comparison uses a different outcome. The primary outcome for the aspirin versus placebo comparison was cardiovascular mortality and for beta-carotene versus placebo it was cancer incidence. Mortality was a leading secondary outcome in both comparisons. There was no interest in the interactions of the two treatments. The aspirin component of the trial was terminated early with a highly significant reduction in fatal and nonfatal myocardial infarction with a trend in cardiovascular mortality. The overall mortality rate was much lower than anticipated in the design so that the PHS was not adequately powered to detect a mortality effect (see Chapter 4) within the planned follow-up time. The beta-carotene arm continued to the scheduled termination and no difference was observed in either cancer incidence or mortality (Hennekens et al. 1996).

Another example of a factorial design was the Alpha-Tocopherol Beta-Carotene (ATBC) trial, conducted in over 29,000 Finnish male smokers.<sup>7</sup> In this trial, two potential prevention agents, alpha-tocopherol and beta-carotene, were evaluated, compared to a matching placebo control. ATBC was also randomized and double-blind. The primary outcome for both beta-carotene versus placebo and alpha-tocopherol versus placebo was lung cancer incidence. Total cancer incidence was a leading secondary outcome. In ATBC neither prevention agent reduced the incidence of lung cancer. In fact, the beta-carotene significantly increased the incidence of lung cancer. In fact, the event repeated by another trial conducted in the U.S. (Omenn et al. 1996).

The Women's Health Initiative (WHI) was a  $2 \times 3$  partial factorial design, evaluating hormone replacement therapy (HRT), a low fat diet, and calcium supplementation in a very large cohort of high risk women.<sup>8</sup> The HRT intervention was compared to placebo in postmenopausal women to test for the reduction in coronary heart disease (nonfatal myocardial infarction and death due to coronary heart disease) and to assess the possible increased risk in breast cancer. The low fat diet was to test for reduction in total cancer incidence and the calcium arm was to evaluate for effect on osteoporosis. The HRT arm had two subcomponents, one that compared estrogen plus progestin with placebo in women with an intact uterus and another component that compared estrogen alone with placebo in women without a uterus. While the primary outcome was coronary heart disease, secondary outcomes included mortality and hip fracture. WHI was a partial factorial because women did not have to participate in all of the three interventions. For the HRT compo-

<sup>6</sup> Steering Committee of the Physicians' Health Study Research Group (1989), Hennekens et al. (1996)

<sup>7</sup> The Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study Group (1994)

<sup>8</sup> Writing Group for the Women's Health Initiative Randomized Controlled Trial (2002), The Women's Health Initiative Steering Committee (2004)

ment comparing estrogen plus progesterin, the trial was terminated early with an adverse effect due to blood clotting and an adverse effect on breast cancer.<sup>9</sup> Osteoporosis was reduced as measured by the hip fracture rate. There was no observed reduction in mortality. The estrogen alone arm was also terminated early due to a significant adverse effect due to blood clotting, with no significant reduction in mortality but with a significant improvement in hip fracture.<sup>10</sup>

#### *Group-randomization Designs*

There are situations in which the intervention of interest cannot be easily administered at the individual level. For example, new sanitation practices in a hospital may prevent certain kinds of infections. It would be difficult to create a large trial to test this hypothesis if practitioners were required to follow different procedures for each patient within a hospital. Instead, it would be easier to treat all patients in a given hospital equally while assigning entire hospitals to different practices. This is the idea behind group-randomized designs. These designs differ from designs randomizing individuals because the groups themselves are not created through the experiment, but rather they arise from relationships among the group members (Murray et al. 2004). This usually induces a correlation among observations from members of each group.

One example of a group-randomized trial was the Seattle 5 a Day study (Beresford et al. 2001). In this study, the intervention consisted of a number of strategies for individual-level and work environment changes. A total of 28 workites in the Seattle area were randomized, half with intervention and half without. The primary outcome was the consumption of fruits and vegetables, as measured by a modified food frequency questionnaire. Other self-reported measures were used as secondary outcomes. These measures were assessed at baseline and at 2 years, using independent cross-sectional samples of 125 workers at each workite. After 2 years, the estimated intervention effect was 0.3 daily servings of fruits and vegetables, which was statistically significant.

A similar trial was known as Teens Eating for Energy and Nutrition at School (TEENS) (Lytle et al. 2006). The intervention was similar to that in the 5 a Day study and consisted of classroom-based curricula, newsletters, and changes in the school food environment in favor of fruits, vegetables, and other nutritious foods. The group units were 16 schools in Minnesota that were randomized to either the intervention or the control arm. Outcome measures included assessments of both the home food environment and the school food environment. The home food environment was assessed with a one time parent survey sent to random subsamples of parents at the end of the study. The results showed that parents of children enrolled in schools receiving the intervention made significantly healthier choices when grocery shopping. Other measures derived from the parent survey did not show significant differences.

The Trial of Activity in Adolescent Girls (TAAAG) (Stevens et al. 2005) also used group randomization. The design called for 36 middle schools to be randomized to control or an intervention with special provisions for opportunities for physical activity. The primary goal was to increase the intensity-weighted minutes of moderate to vigorous physical activity engaged in by girls.

Most group-randomized studies are analyzed using linear mixed-effects models or related approaches. Models of this type are described in Chapter 8 in the context of longitudinal data analysis. The group randomization setting is similar to that of repeated measures data because of the correlation between measurements within randomized units (i.e., schools, workites, etc.). When entire groups are randomized, the notion of “sample size” becomes more complex. Both the number of groups and the number of individuals to be sampled from within those groups must be determined. The necessary calculations are described in Chapter 4.

### **3.3 Non-inferiority Designs**

Historically, most trials have been designed to show that a new intervention is better than or superior to a standard intervention, or that a new intervention added to standard of care is superior to standard of care alone. There are also many situations, however, where a new intervention need not be superior to a standard to be of interest. For example, compared to the standard the new intervention may be less toxic, less expensive, or less invasive and thus have an advantage over the standard, as long as it is not worse than the standard in clinical effect. Many industry sponsors may also want to show that their drug, biological agent, or device is not inferior to a leading competitor product. In cancer or AIDS treatment, a new intervention might have less toxicity and would be of great interest to physicians and patients as long as the effect on recurrence and mortality was almost as good as the standard drug regimen. Trials designed to show that the new intervention is “at least as good as” the control are known as *non-inferiority* trials.

Trials involving diseases that have life-threatening or other severe complications, and for which known effective treatments are available, cannot ethically be placebo controlled if the placebo would be used in place of a treatment known to be effective. (Placebo controls can be used if the new treatment is being used *in addition* to the existing standard.) Thus, unless the new treatment is expected to be superior the current standard, non-inferiority trials must be conducted. On the other hand, the design and conduct of non-inferiority trials can be extremely challenging. For example, in superiority trials, deficiencies in either design or conduct tend to make it more difficult to demonstrate that the new intervention is superior to control, providing incentives that help ensure proper study conduct. For non-inferiority trials, deficiencies can either serve to attenuate treatment differences. In the extreme case, if no subjects in either treatment arm comply with the protocol, the two groups will be indistinguishable and non-inferiority based solely on the formal statistical test will

<sup>9</sup> Writing Group for the Women’s Health Initiative Randomized Controlled Trial (2002)

<sup>10</sup> The Women’s Health Initiative Steering Committee (2004)

be confirmed (although the result will have no scientific credibility). Based on this logic, it is commonly believed that sloppy conduct increases the likelihood of showing non-inferiority (in Chapter 11, however, we show that this is not necessarily the case). Thus, a non-inferiority trial must strive to achieve quality as good as or better than the corresponding superiority trials.

There are also other challenges in non-inferiority trials. One is the choice of the control group. This issue is directly addressed by the ICH-E10 guidance document *Choice of Control Group and Related Issues in Clinical Trials*.<sup>11</sup> In order to demonstrate convincingly that the new intervention is not inferior, the new intervention needs to compete with the best standard available, not the least effective alternative. The selection of the control is not always straightforward. Different alternatives may have different risks and benefits, leading to different levels of compliance. Members of the medical community may not all agree on the treatment to be considered the best standard. One concern that regulatory agencies have is that if the least effective alternative is chosen, then a new intervention might be found to be non-inferior to a less than optimal alternative or control. A series of such non-inferiority trials could lead to acceptance of a very ineffective new intervention. This specific concern leads to an often used requirement that will be discussed later.

Another challenge results from the fact that it is impossible to show statistically that two groups are *not* different. Specifically, any statistic summarizing the difference between groups has an associated variance or confidence interval. To show absolute equivalence requires that the confidence interval have width zero and this would require an infinite sample size. Similarly, to show that one group is strictly non-inferior requires that the confidence interval is strictly to one side of zero (but possibly including zero) in which case we have essentially shown superiority. To overcome this technical difficulty, researchers construct a *zone of non-inferiority* within which the groups are considered *practically* equivalent (see Figure 3.10). The challenge, therefore, is to determine the maximum difference that can be allowed, yet for which the new treatment would still be considered non-inferior. This maximum value is referred to as the *margin of indifference*. The margin of indifference may be expressed as an absolute difference or a relative difference such as relative risk, hazard ratio, or odds ratio. (Mathematically, this distinction is artificial in the sense that relative differences can be expressed as absolute differences on a log scale.) As shown in Figure 3.10, researchers typically use the confidence interval for an intervention effect to draw inferences and conclusions. Figure 3.10 illustrates that if the confidence interval for the treatment difference excludes zero, the new intervention is concluded to be either better (case A) or worse (case D and E). For the non-inferiority trial, researchers want the confidence interval not to exclude zero difference, but rather want the upper limit to be below the margin of indifference (cases B and C). In case B, the estimate of the intervention effect indicates improvement with the upper limit being less

## NON-INFERIORITY DESIGNS

than the margin of indifference. For case C, the intervention effect estimate indicates a slightly worse outcome but the upper limit is still less than the margin of indifference.

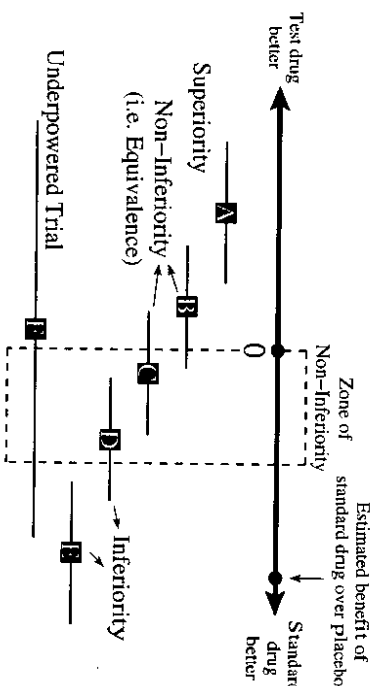


Figure 3.10 Non-inferiority design (absolute difference) (modified from Antman (2001)).

Choosing the value of the margin of indifference is difficult (Gao and Wang *to appear*). The choice must be based on a combination of clinical, ethical, and statistical factors. The size of the margin of indifference depends on the disease and the severity of toxicity or magnitude of the cost or invasiveness relative to the degree of benefit from the standard intervention. For example, researchers may decide that an increase in mortality by 20% may be tolerated if the new intervention had little to no toxicity, or did not require an invasive procedure. Thus, the design would be based on a margin of indifference corresponding to a relative risk of 1.2.

Regulatory agencies often take an approach that imposes two requirements on non-inferiority trials. First, a non-inferiority trial must meet the prespecified margin of indifference,  $\delta$ , when comparing, say, a relative risk of  $RR_{T/C}$  for the new treatment ( $T$ ) to the standard ( $C$ ). Second, the researchers must have relevant data that provides an estimate of the relative risk,  $RR_{CP}$ , of the standard to a placebo ( $P$ ). This estimate is typically based on previous studies, often used to demonstrate the effectiveness of the standard intervention. Then, regulatory agencies may infer the relative risk ( $RR_{TP}$ ) of the new intervention to a placebo control by multiplying the two relative risks,  $RR_{TP} = RR_{T/C}RR_{CP}$ . If, for example non-inferiority is demonstrated versus the control, but the control has only minimal efficacy relative to placebo,  $RR_{TP}$  may show, for example, that despite strong evidence of non-inferiority, the effect against placebo may still be relatively weak. Of course, this interpretation makes a key assumption—that the control relative risk is based on data that is still relevant. This may be difficult to determine as background therapy, patient referral patterns, and diagnostic criteria may change over time.

<sup>11</sup> <http://www.fda.gov/cber/ich/ichguid.htm>

In fact, these arguments are similar to those that were used in section 3.2.1 to indicate why historical control trials have inherent bias.

To make the inference more precise, following Fleming (2007), we transform the relative risks to the log scale, letting  $\beta_{XY} = \log(RR_{XY})$  and  $\sigma_{XY}^2 = \text{Var}(\hat{\beta}_{XY})$ . Using this notation, we assume that the effect of  $T$  relative to  $C$  is given by  $\beta_{TP} = \beta_{TC} + \beta_{CP}$ . Note that since we are assuming that  $C$  is effective,  $\beta_{CP} < 0$  (i.e.,  $RR_{CP} < 1$ ).

Now suppose that we wish to ensure that a fraction,  $p$ , of the effect of  $C$  relative to placebo is preserved:  $\beta_{TP} < p\beta_{CP}$ , or equivalently,

$$\beta_{TC} < p\beta_{CP} - \beta_{CP} = -(1-p)\beta_{CP}. \quad (3.2)$$

Assuming that  $\beta_{CP}$  is known exactly, choose  $\delta = -(1-p)\beta_{CP}$ . Then, non-inferiority of  $T$  relative to  $C$  requires that

$$\hat{\beta}_{TC} + Z_{1-\alpha/2}\sigma_{TC} < -(1-p)\beta_{CP},$$

i.e., the upper limit of the  $1 - \alpha$  confidence interval is below  $\delta$ .

The assumption that  $\beta_{CP}$  is known exactly is unrealistic, so we will need to use an estimate,  $\hat{\beta}_{CP}$ , and also consider its variance  $\sigma_{CP}^2$ , both based on the results of previous trials. Rewriting (3.2), we require  $\beta_{TC} + (1-p)\beta_{CP} < 0$  so that the criterion for showing non-inferiority is

$$\hat{\beta}_{TC} + (1-p)\hat{\beta}_{CP} + Z_{1-\alpha/2}(\sigma_{TC}^2 + (1-p)^2\sigma_{CP}^2)^{1/2} < 0 \quad (3.3)$$

or

$$\hat{\beta}_{TC} + Z_{1-\alpha/2}\sigma_{TC} < -(1-p)\hat{\beta}_{CP} + Z_{1-\alpha/2}[(\sigma_{TC}^2 + (1-p)^2\sigma_{CP}^2)^{1/2} - \sigma_{TC}]. \quad (3.4)$$

so the choice of  $\delta$  is the right-hand side of (3.4). We note, however, that in this case  $\delta$  depends on  $\sigma_{TC}^2$  which may not be known until the trial is completed. When this might be of concern, one may simply prespecify the other parameters,  $p$ ,  $\hat{\beta}_{CP}$ , and  $\sigma_{CP}^2$ , and apply (3.3) once  $\hat{\beta}_{TC}$  and  $\sigma_{TC}^2$  are known.

Certainly, following this approach requires assumptions. One is that the historical estimate of the control effect relative to placebo (e.g.,  $RR_{PC}$ ) is still relevant to the present population and standards of practice. For example, background therapy may have been developed and the patient mix may have changed in important but possibly unknown ways. In addition, the initial trial or trials were based on a particular set of patients or participants who volunteered and they may be different in important respects from the current population being studied. Whether these differences exist or not is usually difficult to determine. This assumption is sometimes referred to as the *constancy* assumption and it may not hold in all cases. Examples exist where two trials conducted consecutively with the same control intervention had significant differences in the control arms between trials (Packer et al. 1996; Carson et al. 2000). In some cases, the historical data may not even exist if the control was established on the basis of an intermediate marker such as lowering blood

pressure, and the next treatment is being evaluated on reducing mortality or morbidity. Second, the percent of the initial control vs placebo effect,  $p$ , to maintain (e.g., 50%) is arbitrary but has a major impact on the choice of  $\delta$ .

Another consideration in setting the value of  $\delta$  is the magnitude of the difference or relative difference that would change clinical practice or patient preference. While the methods described in this section may be of value in providing guidance, other medical considerations should also be considered and adjustments made before the trial design is finalized. Nonetheless, it may be more important to focus on the trial that is being conducted, making sure that the appropriate or best control intervention is being utilized, that the trial is being conducted in the best possible way, and that the new treatment or intervention is compared with the statistically and medically determined value of  $\delta$ . The imputation of the new intervention effect compared to placebo, had a placebo arm been present, while relevant, should not necessarily be the main focus of the interpretation. Control intervention "creep", that is, sequentially choosing weaker and less effective controls, can probably be best prevented by discussions between trial investigators and sponsors with the appropriate regulatory agencies when necessary.

An example of a non-inferiority trial is the OPTIMAAL trial, a trial of a new drug losartan and a standard drug captopril for patients with chronic heart failure (Dickstein et al. 2002). The primary outcome was total mortality. The margin of indifference was determined to be a relative risk of 1.2. Cough is a side effect of captopril use and, as a result, many patients are non-adherent. If the new drug losartan could be shown to be not inferior to captopril, it may be a viable alternative treatment. The standard drug captopril had previously been shown to be superior to placebo with a relative risk,  $RR_{CP}$ , of 0.805 as shown in Table 3.4, the relative risk for the losartan-captopril comparison  $RR_{TC}$ , was 1.12 indicating an increase in mortality for patients treated with the new drug losartan. The imputed relative risk was obtained by multiplying the two relative risks together to get an estimate of 0.906. While this imputed relative risk of 0.906 is favorable, the upper level of the confidence interval 1.26 for the losartan-captopril comparison, did not meet the OPTIMAAL prespecified margin of indifference 1.2. As a result, the new drug losartan was rejected as an alternative to captopril. Of course, the prespecified margin of indifference was a decision based on judgment of the investigators.

Another example that illustrates important issues is two simultaneous trials of a new agent, ximelagatran, compared to the standard control of warfarin. Warfarin is the generally accepted standard to prevent clotting of the blood, but requires frequent monitoring to maintain the correct dose levels and avoid either over- or under-clotting. New agents that could provide the same clinical benefit but without the inconvenience of the intensive monitoring are of great interest. Ximelagatran had been shown in earlier studies to effectively prevent clotting and two trials, SPORTIF III, conducted largely in Europe, and SPORTIF V, conducted largely in the U.S., were designed as non-inferiority trials to assess the ability of ximelagatran to reduce mortality and morbid-

Table 3.4 Results related to the OPTIMAAL trial.

	Rel. Risk	% change
captopril vs. placebo*	0.805	-19.5
losartan vs. captopril†	1.126	12.6
losartan vs. putative placebo	0.906	-9.4 (RR = 0.805 × 1.126)

\* Derived from previous trials  
† Derived from OPTIMAAL

ity.<sup>12</sup> The subjects in both trials suffered from atrial fibrillation, increasing the risk of stroke and thereby the risk of mortality or morbidity. Clearly, warfarin was the appropriate control and the trials were conducted extremely well, achieving excellent clotting control.

Two issues arose, however, that are instructive for non-inferiority trials. The first issue is that the margin of indifference was based on an absolute difference in event rates, assuming an annual event rate of approximately 3%. In fact, the annual event rate observed in the trial was half the assumed rate, raising doubts regarding whether  $\delta$  was in fact too large. The second issue is that there is little reliable data comparing warfarin to placebo, making the imputation of the effect of ximelagatran to placebo problematic. In retrospect, it would have been better to design those trials using a relative scale that would have automatically accounted for the lower than expected event rate. Nevertheless, the lack of historical data regarding the clinical effect of warfarin precludes meaningful imputation of effectiveness relative to placebo. These issues, combined with an observed increase in serious adverse effects related to abnormal liver function, led to the decision by the FDA to not approve ximelagatran based on these trials.

The design and conduct of non-inferiority trials is especially challenging. Many authors have discussed advantages and challenges of non-inferiority trials.<sup>13</sup> While such trials are necessary, the precise design and methods of analysis are still not completely determined and more experience is necessary before such methods will be established.

### 3.4 Screening, Prevention, and Therapeutic Designs

Phase III clinical trials may serve a variety of purposes. Screening or diagnostic trials may evaluate two methods or devices that are designed to detect disease or identify individuals at high risk for disease or an event. Examples of diagnostic procedures include mammography for detecting breast cancer or

<sup>12</sup> Halperin and Executive Steering Committee, SPORTIF III and V Study Investigators (2003), Albers, G. W. on behalf of the SPORTIF Investigators (2004)  
<sup>13</sup> Fleming (2000, 1990), Temple and Ellenberg (2000), Ellenberg and Temple (2000), Hung et al. (2003), Hung et al. (2005)

### SCREENING, PREVENTION, AND THERAPEUTIC DESIGNS

ultrasound for measuring the degree of atherosclerosis in the carotid artery. Prevention trials are designed to intervene in a population at risk for disease but not yet diagnosed, comparing the proposed intervention to standard health care. An example would be giving a cholesterol lowering drug, such as a statin, to individuals identified to have high serum cholesterol, or giving a blood pressure lowering drug to a population with high blood pressure. Therapeutic trials evaluate a new intervention intended to reduce morbidity or risk of death due to diagnosed disease. Using a clot busting drug such as streptokinase in a patient with a heart attack, placing a stent in the coronary arteries for patients with blockage in those arteries, or surgical removal of a tumor in a cancer patient are examples of treatments that could be evaluated in a therapeutic trial to determine if morbidity or mortality would be reduced compared to a patient population not receiving those therapies.

These trials present different design challenges, even if a basic randomized parallel design is utilized. Screening trials such as mammography screening for breast cancer often are very large since, fortunately, the prevalence of breast cancer is low in the general population (Miller et al. 1992). The goal is to identify those individuals in the general population who have early stage breast cancer so that early intervention can be employed. Thus the design must include a strategy for efficiently evaluating a large number of individuals who may have no symptoms or reason to be seeking medical assistance. The false positive rate and false negative rates of the diagnostic procedure are critical in the design of screening trials (Prorok et al. 2000).

Prevention trials typically assess whether a new intervention can reduce the risk of disease developing in disease-free individuals, defined as *primary prevention*,<sup>14</sup> or preventing the recurrence of the disease, defined as *secondary prevention*.<sup>15</sup> In primary prevention trials, individuals at high risk for the incidence of a disease are treated with a drug, device, or procedure to prevent the disease occurrence. For example, in individuals with high cholesterol or high blood pressure, a primary prevention trial using cholesterol or blood pressure lowering drugs assess whether these interventions reduce the incidence of heart attacks or death from the heart attack. Here the individuals entered into the trial may be at higher risk but generally are disease-free and not seeking medical assistance. Thus, recruitment strategies must target otherwise healthy individuals, and convince them to enter the trial. Experience suggests that the yield of patients potentially eligible will likely be no more

<sup>14</sup> Steering Committee of the Physicians' Health Study Research Group (1989), Multiple Risk Factor Intervention Trial Research Group (1982), Writing Group for the Women's Health Initiative Randomized Controlled Trial (2002), Diabetes Control and Complications Trial Research Group (1993), The Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study Group (1994), Hypertension Detection and Follow-up Program Cooperative Group (1979), Lipid Research Clinics Program (1984)  
<sup>15</sup> Beta-Blocker Heart Attack Trial Research Group (1982), MERIT-HF Study Group (1999), Packer (2000), Bristow et al. (2004), ALLHAT Collaborative Research Group (2000), Aspirin Myocardial Infarction Study (AMIS) Research Group (1980), The Coronary Drug Project Research Group (1975), Hulley et al. (1998)

than about 20%, perhaps as low as 5-10%. Therefore, large numbers of individuals must be screened for eligibility and willingness to participate in order to get the desired number of randomized participants.

Since these individuals are not ill, the trial design must account for the likelihood that not all of them will comply with the intervention. That is, the individuals are less likely to take all of their medication, especially if there are undesirable side effects and the sensitivity of the trial to detect a benefit of the intervention will be reduced. In Chapter 4 we discuss sample size calculations that attempt to take into account the anticipated level of compliance. Failure to account for non-compliance can be problematic.

Secondary prevention trials are designed to evaluate whether an intervention can prevent the recurrence of the disease in a population who have had a defining event. For example, we may assess whether drugs such as beta-blockers reduce the occurrence of a second heart attack in a patient surviving a heart attack or whether longer term use of tamoxifen following standard treatment reduces the risk of the breast cancer recurrence. Secondary prevention trials require a different recruitment strategy since these individuals are now identified through an event or occurrence of a disease. Secondary prevention trials, like primary prevention trials, are often large because the recurrence rates may be low. On the other hand, there may be a large number of eligible individuals. While these individuals have been diagnosed with a disease, they may still not fully comply with the intervention being tested. Thus, compliance must be considered in the design of the trial.

For both primary and secondary prevention trials, designing the trial to address the compliance to intervention is critical. As shown in Chapter 4, the sample size increases nonlinearly with the degree of non-compliance. Thus, every attempt must be made in the design to maximize compliance. Characteristics of the participant population may be used to identify individuals likely to have a high degree of compliance. For example, individuals with other competing risks or diseases may have difficulty with compliance. If individuals are not able to come to the clinic for evaluation and intervention support, they may not be able to comply optimally. Once every consideration in the entry criteria and logistical aspects has been covered, then the sample size must be adjusted to account for the remaining degree of non-compliance.

As we will discuss further in Chapter 11, the "intent to treat" principle requires that all participants randomized into the trial must be accounted for in the analysis. Failure to comply with this principle can lead to serious bias in the analysis and interpretation of the results. It is not appropriate to remove participants who do not comply fully with the intervention. Thus, potential non-compliance must be addressed in the design.

Therapeutic trials evaluate new treatments or interventions for patients who are in the acute phase of their disease for the effectiveness in reducing mortality and morbidity.<sup>16</sup> For example, treatments such as clot busting drugs for a

<sup>16</sup> The TIMI Research Group (1988), Volberding et al. (1990), The Global Use of Strategies

patient with an evolving heart attack has proven to be effective in reducing death due to the heart attack. Recruitment strategies for therapeutic trials depend on access to patients and their willingness to participate in the trial. The recruitment pool is generally much smaller than for a primary prevention trial and the yield also much less than 100%, probably also in the neighborhood of 20%.

### 3.5 Adaptive Designs

There are many uses of adaptive methods in the design and conduct of clinical trials. As described earlier, phase I and phase II trials are by their very nature adaptive. In a phase I trial, dose depends on the patient response to the most recent dose. The next stage in a phase II trial does not proceed unless the results from the first stage are favorable. For phase III trials, trial design can be modified as well during the conduct of a trial.

#### 3.5.1 Sequential Designs

A large class of adaptive designs will be described in Chapter 10. Briefly, sequential methods are used to determine when the accumulating results on safety measures and efficacy assessment are so convincing, either favorable or unfavorable to the new intervention, that the trial should be terminated or that the protocol should be modified. The statistical methods are described that also identify when a trial is not going to meet its objectives and continuation is futile.

#### 3.5.2 Outcome Based Adaptive Design

In contrast to the previous section where no data analyses comparing interventions are used for sample size adjustment, *outcome adaptive* trials use the evolving estimate of the intervention effect to make design modifications. Historically, designs such as "play-the-winner" use the success or failure of the last participant to determine the next treatment assignment (see Chapter 5). These designs, however, have not been widely used in clinical trials, despite their superficial appeal.

Other methods have been proposed to modify study design based on the interim estimate of the intervention (Lan and Tost 1997; Fisher 1998; Cui et al. 1999; Shen and Fisher 1999; Chen et al. 2000). One of these is referred to as the weighted Z-statistic method. This method is based on the idea that the type I error rate of the trial is maintained, even if the sample size is changed in response to the observed treatment difference, provided that the

to Open Occluded Coronary Arteries (GUSTO III) Investigators (1997), Moss et al. (1996), Cardiac Arrhythmia Suppression Trial II Investigators (1992), The Diabetic Retinopathy Study Research Group (1978), Fisher et al. (1995), Gruppo Italiano per lo Studio Della Sopravvivenza Nell'Infarto Miocardico (GISSI) (1986)

total information content is unchanged. This is achieved at the conclusion of the trial by decomposing the  $Z$  usual statistic into two components—the  $Z$ -statistic at the time of the design modification and a  $Z$ -statistic derived from the post modification observations.

For simplicity, assume that  $X_i \sim N(\theta, 1)$ ,  $i = 1, \dots, N_0$ , where  $N_0$  is the initial total sample size. Let  $n$  be the sample size at the time the adjustment is made and  $t$  the information fraction,  $t = n/N_0$ . The observed treatment difference is

$$\hat{\theta} = \sum_1^n X_i/n,$$

and the interim  $Z$  statistic for testing  $H_0: \theta = 0$  is

$$Z^{(n)} = \sum_1^n X_i/\sqrt{n}.$$

With no modification the trial would complete with sample size  $N_0$ , and a final test statistic

$$\begin{aligned} Z^{(N_0)} &= \sum_1^{N_0} X_i/\sqrt{N_0} \\ &= \sqrt{\frac{n}{N_0}} \sum_1^n X_i/\sqrt{n} + \sqrt{\frac{N_0-n}{N_0}} \sum_{n+1}^{N_0} X_i/\sqrt{N_0-n} \\ &= \sqrt{t} Z^{(n)} + \sqrt{1-t} \bar{Z}^{(N_0-n)} \end{aligned} \quad (3.5)$$

where  $\bar{Z}^{(N_0-n)}$  is the  $Z$  statistic using only the final  $N_0 - n$  subjects. Note that this a weighted sum of the  $Z$  statistics involving the first  $n$  and final  $N_0 - n$  subjects where the weights are the square roots of the corresponding information fractions.

It can be shown that, provided that the weights remain fixed, we can adjust the final sample size any way that we like, in particular based on  $\hat{\theta}$ , and use the corresponding  $Z$  statistic in place of  $\bar{Z}^{(N_0-n)}$  in (3.5) and still preserve the overall type I error rate.

Suppose that, based on  $\hat{\theta}$ , we propose a revised total sample size  $N$  and the end of the trial we use the test statistic

$$Z^{(N)} = \sqrt{t} Z^{(n)} + \sqrt{1-t} \bar{Z}^{(N-n)}.$$

Then, under an alternative hypothesis  $H_1: \theta = \theta_A$ , unconditionally,

$$EZ^{(N)} = \frac{n + \sqrt{(N_0-n)(N-n)}}{\sqrt{N_0}} \theta_A.$$

Thus, if  $N > N_0$ , we have  $EZ^{(N)} > EZ^{(N_0)}$ , and the power for  $H_1$  will be increased, although the increase will be smaller than if we had used  $N$  as the planned sample size at the start of the trial. Using this approach, the outcomes for the final  $N - n$  subjects are down-weighted (provided  $N > N_0$ ) so that the effective information contributed is the fixed.

## ADAPTIVE DESIGNS

The choice of  $N$  can be subjective, possibly using both prior expectation regarding the likely effect size and the interim estimate,  $\hat{\theta}$ , and its associated variability. Note that since  $Z^{(n)}$  is known, one likely would base the decision on the conditional expectation of  $\bar{Z}^{(N)}$  given  $Z^{(n)}$ .

Tsiatis and Mehta (2003) have criticized this method, suggesting that is not efficient. Others have worried that these trials are prone to bias since speculations about the size of the treatment effect in a blinded trial might arise, based on the size of the sample size adjustment. These more recent response adaptive designs have been used (Franciosa et al. 2002; Taylor et al. 2004) but as yet not widely due to concerns. Jennison and Turnbull (2000) suggest that efficient trials can be designed using common group sequential methods by considering *a priori* a number of possible effect sizes.

Lan and Tost (1997) proposed using conditional power to make an assessment of whether the interim results are sufficiently encouraging to justify a increase in sample size. This concept was further developed by Chen, DeMets and Lan (2000). Conditional power is the probability of obtaining a statistically significant result at the end of the trial, given the current observed trend in the intervention effect compared to control and given an assumption about the true but unknown treatment effect for the remainder of the trial. Conditional power will be described in more detail in the data monitoring chapter (Chapter 10). Let  $Z(t)$  be the normalized statistic observed at information fraction  $t$  in the trial. The conditional power is  $\Pr(Z(1) \geq C|Z(t), \theta)$ , where  $C$  is the critical value at the conclusion of the trial, and  $\theta$  is the standardized treatment difference. The computational details can be found in Section 10.6.3.3. Chen, DeMets, and Lan showed that if the conditional power for the observed trend is greater than 0.50 but less than desired, the sample size can be increased up to 75% with little practical effect on the overall type I error rate and with no serious loss of power.

The methods described above have many attractive features and many control the type I error, as desired. The primary concern with adaptive designs that adjust sample size based on emerging trends is not technical but rather logistical. Trials are designed and conducted to minimize as many sources of bias as possible. If the sample size is adjusted based on emerging trends according to one of the methods described, it is straightforward for those with knowledge of both the size of the adjustment and the adjustment procedure to obtain an estimate of the emerging treatment difference. Of course, following any interim analysis, investigators may have been given clues regarding the current trend—if the recommendation is to continue, investigators might assume that the current results have exceeded neither the boundary for benefit nor the boundary for harm. If there is a formal futility boundary, then additional clues may be given about where the current trend might be. Of course, the difficulty is that data monitoring committees may have reasons to recommend that a trial continue even if a boundary for the primary outcome has been crossed. (Related issues are discussed further in Chapter 10.) Investigators, therefore, may be able to glean knowledge regarding the range

for current trends whether or not the trial is using any adaptive design. For adaptive designs modifying sample size based on current trends, however, the clues may be stronger. The concern is whether this information biases the conduct of the trial in any way. For example, would investigators alter their recruitment or the way participants are cared for? If the trial is double-blind, biases may be less than for non-blinded trials. As of today, there is inadequate experience with these types of adaptive designs to be confident that bias might not be introduced. These designs are still somewhat controversial.

### 3.6 Conclusions

An engineer or architect must finalize the design of a large project before construction of a project begins, for it is difficult, if not impossible, to correct a design flaw once construction has begun. The same is true for the design of a clinical trial. No amount of statistical analysis can correct for a poor or flawed design. In this chapter, we have discussed a variety of designs, all of which have an appropriate application. Each design may need to be modified somewhat to meet the needs of the particular research question. For the phase III clinical trial, the randomized control trial is the most widely used design. While simple, it can be used to address many research questions and is very robust, not relying on many additional assumptions beyond the use of randomization.

A clinical trial should not use a research design that is more complicated than necessary. In general, the practice should be to keep the design as simple as possible to achieve the objectives. In recent years, clinical trials have increasingly utilized the factorial design with success, getting two or more questions answered with only a modest increase in cost and complexity. Greater use of the factorial design has the potential to improve efficiency in cases for which they are appropriate. Statistical methods for adaptive designs that allow for modification of sample size during the course of the trial based on interim trends are available, but there remain concerns regarding their practical implementation. Use of this kind of adaptive design should be considered with caution.

### 3.7 Problems

3.1 Suppose we conduct a phase I trial to find the MTD. We use the daily doses 1g, 1.9g, 2.7g, 3.4g, 4.0g, and 4.6g which correspond to true (unknown) toxicity levels of 0%, 10%, ..., 40%, and 50%, respectively. For this exercise, you may need to use a computer for simulation or, where possible, direct calculation.

- (a) First, consider using the traditional design. The trial begins with the 1g dose and proceeds according to the algorithm described in Section 3.1.1; the last dose without excessive toxicity is the MTD estimate.

- i. What is the expected value of the MTD estimate in this case?
- ii. What is the expected value of the sample size?
- iii. What is the expected number of subjects that would be treated at or over the 40% toxicity level (i.e., 4g) in this case?
- iv. What modifications could you make in the design so that the expected value of the toxicity of the MTD estimate will be approximately  $1/3$ ?

(b) Next, consider the modified design (allowing steps down in dose), starting at the third level (2.7g), stopping after 15 patients.

- i. What is the expected value of the MTD estimate (use the stopping dose to estimate the MTD)?
- ii. What is the expected number of patients who would be treated at or over the 40% toxicity level?

(c) Finally, consider using another design with single-patient cohorts. Start at the first dose level, increasing the dose for the next subject when a non-toxic response is observed. When a toxic response is observed, a second stage begins at the previous (lower) dose level. If two subsequent non-toxic responses are observed, the dose increases. If a toxic response is observed, the dose is reduced. Stop after a total of 15 patients.

- i. What is the expected value of the MTD estimate (again, use the stopping dose to estimate the MTD)?
- ii. What is the expected number of patients who would be treated at or over the 40% toxicity level?

(d) How would you derive confidence intervals for the MTD using data from the three designs above? Demonstrate your confidence intervals' coverage properties via simulation.

3.2 Consider Fleming's three-hypothesis formulation for phase II trials with binary outcomes. Suppose you have a single arm trial where  $\pi_1 = 0.05$ ,  $\pi_2 = 0.25$ , and the total sample size is 30. Set the level for  $H_0$  and  $H_2$  each to 0.05, i.e.,

$$P(\text{reject } H_0 | \pi_T = 0.05) \leq 0.05 \quad \text{and} \\ P(\text{reject } H_2 | \pi_T = 0.25) \leq 0.05$$

(a) Give the three critical regions that would be used at the trial's conclusion. State the probabilities of observing an outcome in each of the three regions, given that  $H_2$  is true ( $\pi_T = 0.3$ ).

(b) Suppose that, after fifteen patients, only one success is observed. What is the probability, conditional on the interim results, of rejecting  $H_0$ , given that  $H_2$  is true ( $\pi_T = 0.3$ )? What is the probability, conditional on the interim results, of rejecting  $H_0$ , given that  $H_0$  is true ( $\pi_T = 0.05$ )?

(c) Before the trial starts, suppose it is determined that another identical trial, using the same critical regions, will be performed following this one in the event that no hypothesis is rejected. What is the probability, given  $H_0$  is true ( $\pi_T = 0.05$ ), that  $H_0$  will be rejected? What is the probability, given  $H_2$  is true ( $\pi_T = 0.25$ ), that  $H_2$  will be rejected? Should the critical regions be adjusted to conserve the overall error rates? Explain.

3.3 Consider a non-inferiority study in which we are interested in the probabilities of failure  $\pi_C$  and  $\pi_E$  in the control and experimental arms, respectively. In particular, we are interested in the following two pairs of hypotheses: absolute (additive) non-inferiority,

$$\begin{aligned} H_0 &: \pi_E - \pi_C \geq 0.1 \\ H_1 &: \pi_E - \pi_C < 0.1; \end{aligned}$$

and relative (multiplicative) non-inferiority,

$$\begin{aligned} H_0 &: \frac{\pi_E}{\pi_C} \geq 2 \\ H_1 &: \frac{\pi_E}{\pi_C} < 2. \end{aligned}$$

There are 80 subjects in each arm. Assume the sample is large enough to use the normal approximation.

- What test statistic do you suggest for the hypotheses of absolute non-inferiority?
- For a one-sided test at the 0.05 level, show the rejection region on a plot with axes corresponding to the observed failure proportions in the two arms.
- What test statistic do you suggest for the hypotheses of relative non-inferiority?
- For a one-sided test at the 0.05 level, show the rejection region on a plot as in the previous case.
- Which test do you think will be more likely to reject  $H_0$  when  $\pi_E = \pi_C = 0.1$ ?
- In general, supposing that  $\pi_E = \pi_C = 0.1$ , for what values (approximate) of these probabilities, if any, do you think that your absolute non-inferiority test will be more likely to reject  $H_0$ ? For what values, if any, do you think that your relative non-inferiority test will be more likely to reject  $H_0$ ?