7

RANDOMIZED CLINICAL TRIALS

This chapter and the next give illustrations and applications of the material presented in earlier chapters. In this chapter, we discuss the design and analysis of clinical trials. We focus on the themes of the earlier chapters: randomization, hypothesis testing and sample size, and estimation and analysis. We also discuss some unique aspects of clinical trials dealing with ethical issues, complexity, and regulatory oversight. These issues are illustrated by four clinical trials. We can only highlight some of the statistical issues; whole books have been written on these topics. At the end of this chapter, we give some references.

A clinical trial is an experiment to assess the *efficacy* and *safety* of two or more treatments. The word *treatment* in this context is any therapeutic intervention; it could be a biological product such as platelets, a drug such as a statin, an appliance such as an artificial hip, or a behavioral intervention. By efficacy is meant a clinically meaningful effect (endpoint), that is, an effect that is of tangible benefit to the patient. Safety refers to the absence or tolerable presence of side effects. There always is a trade-off between efficacy and safety, with the result that the final assessment of treatment is in terms of *benefits* and *risks*. For example, there are no cures for multiple sclerosis. Until 2010, palliative treatments involved injections; now there is a new palliative treatment consisting of tablets, a major convenience to the patient. However, the new treatment has the potential of causing optical problems (macular edema); hence, there is the question of comparing benefit with risk.

Most clinical trials share the following design characteristics.

- The design of a clinical trial is usually fairly simple: a completely randomized experiment (*parallel groups design*) for comparing treatment. Reasons include

Design and Analysis of Experiments in the Health Sciences, Gerald van Belle and Kathleen F. Kerr.

large variability among human subjects, ethical and cost constraints, sequential enrollment of subjects, the inappropriateness of testing many treatments on the same subjects, and the constant risk of dropout. There usually is some kind of balancing of assignments over time since subjects are recruited sequentially. For example, three treatments are assigned in blocks of nine so that there is balance after every nine subjects. (Usually, the size of the block is hidden from the investigator to maintain blinding.) What makes the balancing challenging is that, as is commonly the case, multiple centers are involved with the need to this balance within each center.

While the completely randomized design is the most common, we will discuss alternative designs such as Latin squares, sometimes used in dentistry where different areas of the teeth or jaw may receive specific treatments. On the whole, there are few clinical trials that use designs more complicated than those discussed in this text.

- A clinical trial involves longitudinal data. Hence, all the issues discussed with repeated measures designs crop up. Minimizing the dropout rate and missing data is a paramount objective. This concern is especially acute in clinical trials related in part to the length of such trials. It is not uncommon for a clinical trial to run 3 or 4 years. Some are even longer: a study to assess the effect of an 8-year, school-based intervention on smoking behavior at grade 12 (and 2 years post-high school) took more than 12 years (see HSPP trial in Table 7.1).
- Multicenter trials are scientifically and logistically complex; it is not uncommon to have five or more centers participate. Any time there is more than one center issue such as standardization of protocols, data collection, and reporting crop up. Other level of complication involves regulatory oversight from local institutional review boards to government agencies. One of the most complex clinical trial structures is that of a network of centers (sometimes called a *consortium*) that carries out multiple clinical trials. While such networks exist to create efficiency and continuity, they run the risk of suffocation by committee; it is not uncommon to have separate committees to deal with protocol review, data and safety monitoring, management, study monitoring, publications, and on top of the organizational food chain, the Executive Committee.
- There are inherent ethical requirements for a clinical trial. The principal investigator has primary responsibility. There are government regulations that require the principal investigator to follow specific procedures. Most trials require review and approval by a local review board known as an Institutional Review Board (IRB) in the United States, Research Ethics Board (REB) in Canada, and some type of ethics committee in the European Union. These bodies exist to protect human subjects. Ethical requirements permeate the design of clinical trials. (An especially knotty ethical problem occurs in emergency medicine where time is of the essence and neither subjects nor family members can give *informed* consent. Problem 4 at the end of the chapter deals with this topic.)
- Given the complexity of clinical trials and their long duration, they are expensive not only in terms of money but also in terms of effort and commitment of all

the senior—and not so senior—investigators. In terms of expense, costs usually run into the millions of dollars. The cost is borne either by government or by a pharmaceutical firm; it is not uncommon to have a cost-sharing arrangement. With a pharmaceutical or device firm involved, there also arise the issues of who owns the data, to what extent can the firm influence publications of papers, the specter of conflict of interest, and so on.

- The randomized clinical trial (RCT) has reached an exalted status in health-related studies. The RCT grounds what is known as *evidence-based medicine* (EBM) that stresses that validity of a medical procedure can only be anchored securely in the results from a properly conducted randomized controlled clinical trial. A "cottage industry" has emerged for evaluating such trials. The most prestigious and widely cited are the reports from the Cochrane Collaboration (www.cochrane.org) that reports routinely on clinical trials.
- Most clinical trials are registered. The most common registry is one maintained by the government of the United States. The registry is open to the public (http://www.clinicaltrials.gov/). Each trial gets a unique identifier. At end of 2011, the registry had information on more than 100,000 trials from 180 countries. Each trial is identified by a unique number of the form NCT followed by eight digits. You should become very familiar with this source since it contains information about the design of the trial, eligibility and exclusion criteria, status of the trial, and other information such as a list of publications coming from the trial.

There are two broad classes of clinical trials: *public health clinical trials* and *pharmaceutical clinical trials*. In the former category are studies aimed to affect public health practice, for example, assessing effective ways to reduce smoking among teenagers or the effect of diet on cardiovascular disease. Such trials are primarily funded by government agencies in response to a perceived public health need.

In the latter class are trials dealing with the evaluation of new medicinal preparations or appliances. One characteristic of these trials is that they are of much shorter duration because a rapid clinical endpoint is considered and are frequently sponsored by a pharmaceutical firm with the intention of getting approval from a regulatory agency to market its product. The distinction between these two types is not absolute but is useful. As mentioned, it is not uncommon to have collaboration between industry and government. Government can share in the burden of testing the efficacy of *orphan drugs* that have a limited commercial market; industry often supplies, free of charge, the medications for public health trials. Both kinds of trials rely heavily on the design and analysis of experiments.

7.1 ENDPOINTS

In Chapter 1, we discussed the four fundamental questions of scientific research. The first, "what is the question?"; the second, "is it measurable?"

Regarding the first, the clinical trial requires a result, or endpoint, that has immediate clinical relevance during the trial and afterward. That is, it should reflect

tangible benefit to the patient. This clearly includes disease outcomes, especially cancer and cardiovascular disease and, as the population ages, dementing illnesses. For example, in diabetes it may be the prevention of amputation of limbs or stroke. The second requirement is that the endpoint be measurable. In many instances, this will be straightforward, such as amputation of a limb in diabetes research, or survival in the case of pancreatic cancer. The requirement becomes more challenging when diseases such as chronic depression are considered. What is a clinically meaningful endpoint in this case? And how will it be measured? In all these cases, the end(point) justifies the means.

In many trials, the key objective is the prevention or delay of death. But it may not be feasible to wait until this event occurs, so some intermediate endpoint is sometimes chosen. The choice of such *surrogate endpoints* is very challenging. It requires knowledge of the disease process. An example of a surrogate endpoint is blood pressure. It is known that high blood pressure is associated with the risk of stroke, so it would seem obvious that reducing blood pressure is a "good thing." Another example is tumor shrinkage in the case of cancer. On the whole, researchers and regulatory agencies take a dim view of surrogates, since not enough is known about the disease process and how change in the surrogate will affect the clinical endpoint. The challenge is that a surrogate correlated with a clinical endpoint is not necessarily a good surrogate for the following reasons: the surrogate is not in the causal pathways of the disease process but associated with the outcome, for example, PSA (prostate-specific antigen) and prostate cancer. A second reason is that there may be multiple causal pathways to a clinical endpoint. For example, there are (at least) three considerations in the link between diabetes and outcome (such as amputation of an extremity): smoking, diet, and control of sugar level. If the treatment addresses, say, control of the sugar level, there may not be any effect on the outcome because the other two factors are not controlled. And it may well be that the agent that controls the sugar level has serious side effects. Third, an intervention may bypass the causal pathway. Finally, treatment of surrogate endpoint may actually be harmful to the patient (see van Belle et al. (2004) for examples).

There are two criteria for a valid surrogate. first, it must be correlated with the clinical endpoint (a necessary, but not sufficient, condition). Second, the surrogate "must fully capture the net effect of the intervention on the clinical endpoint" (Prentice, 1989). The second condition is very difficult to verify in practice.

PSA and similar measures are known as *biomarkers*. New biomarkers are announced almost daily with promise of a key component in the treatment of disease. Ioannidis and Panagiotou (2011) indicate that the initial enthusiasm for a biomarker is subsequently dampened—this may represent regression to the mean. The slogan "from the bench to the bedside" in the case of biomarkers becomes "from the bench to the bedpan."

7.2 RANDOMIZATION

As discussed earlier in this text, randomization is a key to the validity of clinical trials. Randomization provides the best assurance of comparability of groups and provides

the basis for the statistical analysis. Given its importance, randomization needs to be very carefully described in the protocol and strictly adhered to. The randomization may be blocked or stratified by institutions or other participating units.

Another feature of randomization in the RCT is that it does not need to be at the individual level but could be at the level of a *group*. The Peterson et al. (2000) study—see below—involved randomization of school districts.

In large-scale clinical trials, enrollment may take several years so that the assignment to treatment is an ongoing process, usually computerized to be efficient and to satisfy certain allocation restrictions. For example, randomization to treatments may be in blocks of specified size in order to maintain balance among treatments and over time. (Typically, the block size is confidential, so neither participants nor sponsors can guess the next allocation.) An automated computerized allocation of subjects has the advantage that treatment assignments can be made any time during the day or week. In older clinical trials, treatment assignment required contact with coordinating center personnel who typically were available only during regular office hours. It is desirable to have treatment assignment as close as possible in time to treatment initiation.

7.3 HYPOTHESES AND SAMPLE SIZE

The inferential framework for clinical trials is identical to the approaches described earlier. Null and alternative hypotheses are specified, within-treatment variation is estimated, and treatment effects postulated based on pilot studies or other sources. Given this frame work, sample sizes can be calculated. The recommended approach, as before, is to use the hypothesis testing framework in the design of the study and confidence interval approach for the analysis.

Some unique characteristics of clinical trials are (1) given the length of the trial an interim analysis may be desirable, (2) since human subjects are involved there is an ethical imperative to stop the study as soon as possible if unanticipated and unacceptable adverse events are observed, and (3) the usual hypothesis testing framework may be extended to include situations of equivalence or noninferiority of treatments when compared to standard treatments.

Often, large-scale clinical trials have several key endpoints necessarily and sample size calculations that are based on a confluence of considerations. This makes it difficult at times to determine why a specific sample size is ultimately selected. All RCTs maintain a prespecified Type I error rate, usually 0.05 for the overall study.

Interim analyses, a characteristic of many RCTs, are planned in detail to maintain the overall Type I error rate. For example, the Casa Pia study, discussed below, planned interim analyses every year using a specified amount of the Type I error so that the total added up to 0.05. For the first year of the study, the interim analysis used 0.0125; for years 2–6, 0.0015; and for year 7, the remaining Type I error of 0.030. Why the strict adherence to the Type I error rate? To prevent unwarranted conclusions of treatment effectiveness. A regulatory agency does not want to approve a treatment that is not effective—it's easier to bar a product from the market than to remove one subsequently shown to be ineffective.

Sample sizes may be very large because the occurrence of clinically important events is relatively rare, for example, death. This also illustrates a key sample size issue: it is the number of events rather than the number of subjects that drives the sample sizes.

In addition, sample sizes are increased to compensate for refusal to participate and dropouts. Usually, the sample size is calculated and then inflated by the estimated proportions of refusals and dropouts. For example, the SELECT study discussed in the next section has as primary endpoint the clinical incidence of prostate cancer. The study that will last 12 years worked with an estimate base rate of prostate cancer of 6.6% after 12 years. The study assumes a 25% reduction in the incidence associated with one of the treatments. Just on this basis, the sample size assuming 80% power and the usual binomial model would lead to an estimated sample size of 4000 men per group. But factors that drive the sample size higher are fewer than 12 years of observations on a large fraction of the sample, dropouts, five prespecified comparisons, a higher power (95%), and other considerations. So the estimated sample size of 32,000 men (16,000 for the two main comparison groups) is understandable.

7.4 FOLLOW-UP

A clinical trial usually involves follow-up of subjects with the potential risk of dropout. It is imperative that a high rate of follow-up of the endpoint be achieved. The converse is that there must be low *attrition*: dropouts may introduce bias (not missing at random); especially if the rates differ by treatment. The high rate is required in order to maintain the validity of the randomization—dropouts destroy randomization. The design of the trial should include strategies (and provision of funds) for ensuring that follow-up is successful. For example, at the time of enrollment, a subject provides names and addresses of next of kin, or neighbors to the investigators.

Attrition during a clinical trial is a fact of life but the investigator needs to assure that it is minimized through adequate follow-up. It will not do to base sample size calculations on, say, a 50% attrition rate when 20% is achievable with little additional effort. Regulatory agencies are very critical of randomized clinical trials with attrition rates greater than 20%.

In the next section, we discuss some statistical approaches to dealing with attrition—second best to achieving a low attrition rate.

7.5 ESTIMATION AND ANALYSIS

Estimation and analysis are carefully prespecified in the protocol of a clinical trial based on *primary* and *secondary* endpoints. The primary endpoints are the key elements in the hypothesis–sample size–conduct–analysis chain. The primary endpoint is the basis for satisfying the requirements of efficacy and labeling in the pharmaceutical trial.

All clinical trials have to deal with potential crossovers and dropouts. Crossover occurs when a subject is randomized to one treatment but receives another treatment.

Reasons can vary from the protocol not being followed to patient choice to switch to another therapy; for example, a prostate cancer subject assigned to radiation therapy decides to have surgery. The two most common approaches to crossover are *intent to treat* (ITT) or *treatment received* (TR), also called *per-protocol* analysis, with ITT the default standard. In an ITT analysis, subjects are classified by the treatment assignment at randomization. This kind of analysis tends to be conservative but is considered the most robust and less subject to bias. One way to minimize the issue is to carry out the randomization as closely as possible to the treatment being given. In many clinical trials, a pool of eligible subjects is created but randomization is not carried out until the treatment has to be selected. This is a good principle of design but may run afoul of subject, and clinician, anxiety. For example, suppose the alternatives are radiation or chemotherapy after "watchful waiting" for disease progression. In this situation, it may be very difficult to wait until the last moment. There are situations where the ITT approach may be questioned; see Piantadosi (2005) for a very useful discussion and references.

Dropouts are a challenge since there is no endpoint. There are several standard strategies. One is to impute a value based on matching the dropout characteristics with a subject who has not dropped out. Repeating this process several times leads to *multiple imputation*. Another strategy is to use the last outcome value observed (LOCF, see Section 6.6.2). These kinds of considerations have led to the development of a large body of statistical methodology on how to deal with subjects who drop out some time during the trial and there is a huge literature discussing these strategies. Cook and DeMets (2008) is a good place to start. It must be emphasized that none of the above approaches can overcome deficiencies due to crossover and dropouts. Public health clinical trials tend to focus on a specific main endpoint with intermediate values of secondary interest only reflecting an interest in *effectiveness* rather than *efficacy*. Pharmaceutical clinical trials do use intermediate points extensively.

Once the appropriate endpoints and data for statistical analysis have been created, the actual analysis is fairly straightforward—in part because the designs are basically simple. Roughly speaking, there are two types of endpoints: binary and measurement. Binary endpoints could be success or failure. Another binary endpoint is survival status and an associated measurement variable, length of survival. Binary endpoints, other than survival, are commonly handled by logistic regression. Survival endpoints are handled by survival analysis. Measurement variables are most commonly analyzed by methods discussed in this book. In the examples in the next section, many of these approaches are used. Again, the book by Cook and DeMets (2008) is a good place to start learning about these methods.

7.6 EXAMPLES

Rather than giving one example, we briefly look at four clinical trials (RCTs) to illustrate unique features and similarities. These RCTs varied in their treatment structure, subjects, randomization, and endpoints. Table 7.1 highlights different aspects of these trials.

We note some common features, indicating a standard approach to these clinical trials. Each of these studies was a response to a *Request for Proposals* (RFP) or *Request for Applications* (RFA) by the National Institutes of Health of the United States. These requests are based on an identified need to study a particular area of health care and money is set aside to fund these studies. Competition for getting an award is intense. Proposals are reviewed by an independent scientific body that not only ranks the applications and applicants but also judges whether any application meets scientific standards.

All four trials involved some kind of randomization. In terms of sample size calculations, all assumed a Type I error rate of $\alpha = 0.05$. The power is typically higher than the default 80% of sample size formulas such as the one in this book. The reason is that given the huge expense there is pressure to make sure that a treatment effect, if present, will be detected.

All of these proposals were reviewed and approved by one or more Institutional Review Boards; in the case of multicenter studies, each center has its own board. Getting consensus among these boards is not always easy and is always time consuming. Other common features include *Data Safety Monitoring Boards* that monitor the study as it proceeds, an extraordinary time commitment by the key investigators and their staffs, and huge costs.

The trials summarized in Table 7.1 lasted many years beyond their initial design life span. One reason is that often additional follow-up to longer term endpoints can yield important new information. These large-scale public health trials are like old soldiers, they don't die, they just fade away with *coup de grâce* administered by the funding agency when funding ceases—although some studies are creative in finding other sources of funding. The payoff from these trials is an opportunity to modify fundamentally health delivery practice. This is of interest not only to researchers but to sponsoring agencies as well, since a great deal of medical expenses are paid out of the public purse—for example, medical benefits for retirees. It also makes good politics to sponsor efforts to improve a nation's health.

We discuss each of the four examples briefly and illustrate common and unique features.

1. Casa Pia Study of the Health Effects of Dental Amalgam in Children

Dental amalgam, widely used in dentistry, contains elemental mercury that emits a small amount of mercury vapor that is a known neurotoxic agent. An alternative is a resin composite that does not contain mercury but has the disadvantage of not lasting as long. A study to assess the amalgam's effect on neurobehavioral and neurological outcomes was carried out in Lisbon, Portugal, among students of the Casa Pia school system. Students were randomized to either amalgam-based dental restoration or resin composite materials. A total of 507 children, aged 8–10 years, were randomized with 253 in the amalgam group and 254 in the resin composite group. This study that lasted 7 years was not able to detect statistically significant differences between the amalgam and composite groups in the specific neurobehavioral neurological outcomes. There were no borderline significant results in these outcomes.

Table 7.1 **Comparison of four public health randomized clinical trials.**

	Clinical Trial			
	Casa Pia	HSPP	WHI	SELECT
Title		Hutchinson Smoking Prevention Project	Women's Health Initiative	Selenium and Vitamin E Clinical Trial
Objective	Safety of dental amalgam (which contains mercury)	In-class curriculum to prevent smoking in teenage years	Hormone replacement therapy (HT) and diet effect on heart disease, cancer, and fractures	Selenium and vitamin E effect in preventing prostate cancer
Date of RFP	1993	1983	1991	1999
Subjects	Children between the ages of 8 and 1	Children in grade 3	Women under the age of 60	Men over the age of 50
Treatment structure	Amalgam or resin composite	Classroom-based social influences curriculum, no intervention	Hormone replacement therapy, diet, calcium supplementation	None, selenium, vitamin E, or both
Design structure	Completely randomized (CRE)	Randomization within 20 pairs of school districts	Partial factorial	Factorial
Primary outcome(s)	Neurobehavioral, nerve conduction velocity	Smoking status in grade 12 and 2 years post-high school	Cardiovascular disease	Prostate cancer
Sample size	507 children	40 school districts, 8388 children	93,676; CRE + observational	32,400
Power—see text	>97%	97%	86%	89%
Blinding	Psychometrists blinded	Evaluators blinded	Yes	Yes
Location	Portugal	Washington State	United States—40 clinical centers	United States–300 clinical centers
Intervention start and end dates	1997–2005	1984–1999	1993–2002	2001–2008
Analysis	O'Brien test and Hotelling *T* test	Permutation test among the 20 pairs	Survival analysis	Survival analysis
Conclusion	No statistically significant differences	No statistically significant differences	HT arm terminated early due to significant side effects	Terminated early; no tx effect
Registration #	NCT00066118	NCT00115869	NCT00000611	NCT00006392

See text for further elaboration and for explanation of acronyms.

This study is interesting in that although the alternative hypothesis was two-sided, the significance levels were divided unequally with the overall tests for the adverse effect of amalgam set at 0.03 and resin composite at 0.01 (another 0.01 was used for the Hotelling T test). The O'Brien test takes into consideration that there were multiple outcomes. In fact, the O'Brien test was modified to take into account repeated measures—a good example of how each trial presents unique statistical challenges (Leroux et al., 2005).

Although this study was a safety study, not an efficacy study, it noted that starting at "5 years after treatment, the need for additional restorative treatment was approximately 50% higher in the resin composite group" (DeRouen et al., 2006).

2. Hutchinson Smoking Prevention Project (HSPP)

This trial "aimed to attain the most rigorous randomized trial possible to determine the long-term impact of a theory-based, social-influences, grade 3–12 intervention on smoking prevalence among youth" (Peterson et al., 2000). Randomization was at the school district level, with 20 pairs of school districts randomized to either a school-based tobacco prevention program or control. The schools were paired on the basis of (1) tobacco use in older students (teens) determined immediately after recruitment of the school district and (2) location (east or west of the Cascade mountains). This trial is an example of group randomization rather than individual randomization.

In this large trial, participants were 4177 third graders in the 20 experimental school districts and 4211 third graders in the 20 control school districts. See also Figure 7.3. No statistically significant differences were found in the prevalence of daily smoking either at grade 12 or 2 years after high school. The study concluded that "consistent with previous trials, there is no evidence from this trial that a school-based social-influences approach is effective in the long-term deterrence of smoking among youth" (Peterson et al., 2000). In an accompanying editorial, Clayton et al. (2000) asserted that this study "suggests that the social cognitive learning approach . . . may be virtually useless in explaining what causes some people to smoke and others not to smoke.. . . "

The group randomization permutation procedures used for the analysis accommodated the correlation of responses among children from the same school district. The test is nonparametric, that is, does not require modeling or distributional assumptions, based solely on the permutations of outcomes among the 20 pairs of schools. Three possible effect modifiers of interest were identified at the start of the study: a child and family risk of smoking (a person/family variable), school enrollment (as an exposure variable), and school risk of smoking (a school/environment variable).

In a subsequent study, equally carefully carried out, Peterson et al. (2009) showed that a personalized telephone counseling intervention for youth smoking *cessation* was effective.

3. Women's Health Initiative (WHI)

The Women's Health Initiative was and is one of the largest clinical trials undertaken in the United States and perhaps the world. The trial started in 1992 and was

slated to continue until 2010, with every prospect of continuation beyond that time. It ultimately involved 93,676 postmenopausal women in the age range 50–79. Enrollment was started in 1993 and concluded in 1998. The primary aim was to evaluate the health benefits and risks of four interventions: dietary modification, two types of postmenopausal hormone replacement therapy, and diet supplementation of calcium and vitamin D. The design was a "partial factorial" with women with an intact uterus receiving one type of hormone replacement therapy and those who had a hysterectomy prior to randomization another form.

Endpoints included the occurrence of breast cancer, cardiovascular disease, stroke, colorectal cancer, and hip fracture.

Both hormone-related treatments were stopped early in 2002 when it became clear that the risks exceeded the benefits: increases in breast cancer, cardiovascular disease, and stroke; and decreases in hip fracture and colorectal cancer (Writing Group for the Women's Health Initiative Investigators, 2002).

A good starting point for reading about this study is The Women's Health Initiative Study Group (1997). A discussion of statistical issues can be found in Prentice et al. (2005).

4. Selenium and Vitamin E Cancer Prevent Trial (SELECT)

Prostate cancer is a leading cause of cancer death in males—but is relatively rare. This trial, a prevention trial, investigates the effects of selenium, vitamin E, or both on the incidence of prostate cancer in males. The design is a 2×2 factorial (vitamin E only, selenium only, vitamin E and selenium, or none). Since vitamin supplements contain these ingredients, the study supplies participants supplements with these items included only in the appropriate groups. A total of 32,400 men have been randomized (8100 per treatment group). One reason for the large number is the low incidence of prostate cancer. In Year 7 of the study (2008), the independent Data and Safety Monitoring Committee found that the treatments alone or together did not prevent prostate cancer. It also determined that it was very unlikely that the selenium and vitamin E supplementation would ever produce a 25% reduction in prostate cancer as the study was designed to show. As a result, participants were told to stop taking the supplements as part of their participation in the trial. Since the preparations are available over the counter, the investigators could not control participants continuing to take these medications (even though they may have been on placebo or only one treatment during the trial). The results of the trial were reported in 2009 by Lippman et al. (2009). Follow-up is continuing.

7.7 DISCUSSION AND EXTENSIONS

7.7.1 Statistical Significance and Clinical Importance

There are statistical and clinical aspects to the outcome of a trial. Figure 7.1 provides a schematic. The context is a two-group parallel study comparing a test treatment

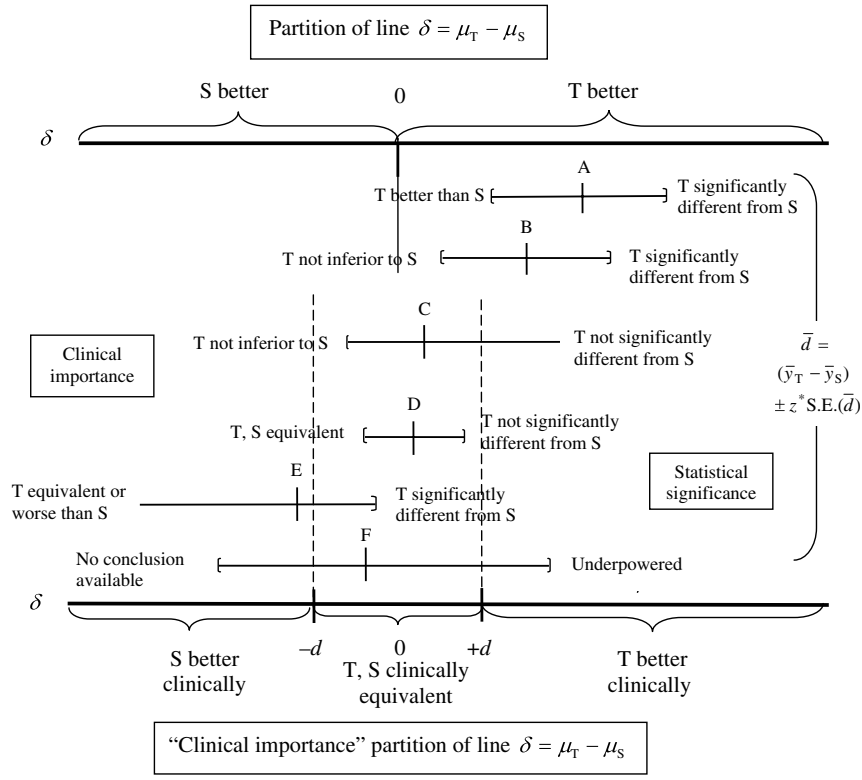

Figure 7.1 Partition of the line $\delta = \mu_T - \mu_S$ and its clinical importance for a two parallel group clinical trial comparing a test preparation (T) with a standard preparation (S). Conclusions are indicated for different results with point estimates and confidence intervals. The confidence coefficient z^* is chosen to generate a two-sided (Cases A, B, D, and F) or one-sided confidence interval (Cases C and E). See text for elaboration.

(T) with a standard treatment (S). The population means of the two treatments are denoted by μ_T and μ_S with $\delta = \mu_T - \mu_S$. The quantity δ is estimated by the sample mean difference $\bar{d} = \bar{y}_T - \bar{y}_S$, where $\bar{y}_T$ is the sample mean for treatment T and $\bar{y}_S$ the sample mean for treatment S. We can construct a two-sided $100(1 - \alpha)$ confidence interval with the usual interpretation that if the interval does not straddle 0 the null hypothesis of $\delta = 0$ is rejected. Case A and Case C (ignoring the one-sided arrow for the time being) illustrate this discussion.

The clinician may be interested in a more refined assessment, as illustrated by the lower half of Figure 7.1. Specifically, there is a region $-d$, $+d$ where the treatment differences are small and not clinically relevant. In this region, the treatments are considered equivalent. (The choice of d is crucial and will be discussed below.) The outcome space is then divided into three regions: S better clinically, S and T equivalent, T better clinically. The lower half of Figure 7.1 illustrates this partition. In the region

$-d \geq \delta \leq +d$, the treatments are clinically equivalent. This leads to three possible interpretations of the results of a clinical trial:

- *Nonequivalence:* confidence interval completely outside $-d$, $+d$, (Case A).
- *Equivalence:* confidence interval completely inside $-d$, $+d$ (Case D).
- *Noninferiority:* lower bound of confidence interval $> -d$ (Cases B and C).

The region of equivalence is of great interest to pharmaceutical firms who may want to develop a generic drug equivalent to a standard drug. Or, alternatively, a new formulation with fewer side effects is being considered by researchers. The data for the inference are based on a confidence interval based on the observed difference $\bar{d}$ and its standard error S.E.($\bar{d}$).

The arrows in Figure 7.1 are either bidirectional or unidirectional. What is going on? If we emphasized the clinical importance and, for example, wanted to show that treatment T is not inferior to a standard treatment, we would construct a one-sided confidence interval (Case C). This has implications for sample sizes and power. Julious (2004) derives the appropriate critical values of α and β for generating sample sizes and confidence intervals.

How to pick the value d? There are many rules—suggesting that there are no "hard and fast" rules. One rule is to postulate that the mean for the test treatment does not differ by more than, say, 10% from the standard treatment. Another rule is to specify a clinically meaningful difference and then pick d to be half of that value.

Sample size calculations for the equivalence and noninferiority situations are complicated by the following: (1) two alternative hypotheses are tested corresponding to the bounds of the equivalence interval and (2) uncertainty about the value of the parameter δ. The two alternative hypotheses require partitioning of the Type II error β. Uncertainty about δ may result in substantial increases in sample size. If, for example, $\delta = 0.25d$, the sample size may increase by about 50%. See Julious (2004) for details. da Silva et al. (2009) contains a very readable account about inferiority and noninferiority testing. If you are involved with a clinical trial dealing with equivalence or noninferiority issues, it may be wise to consult a biostatistician.

7.7.2 Ethics

The principal investigator of a clinical trial is responsible for the ethical conduct of a clinical trial ensuring that the study is being conducted in accordance with regulatory guidelines for the protection of human subjects.

In a perfect world of science, politics, and values, there would be no need for checks and balances in research using human subjects. Unfortunately, this is not the case and institutional review of proposed research is now the standard. The committees, or boards, that carry out this responsibility are primarily concerned with the research design of a proposed study, the consent process, and, more recently, the collection of confidential information. A typical charge to an IRB is "approval should occur only when the Committee agrees that the project has scientific merit, a reasonable

risk/benefit ratio, equitable subject selection, adequate privacy and confidentiality protections, and, unless waived, informed consent procedures are adequate" (Human Subjects Review Committee, Group Health Cooperative, Seattle, Washington).

In the United States, three ethical principles guide the use of human subjects: (1) respect for persons or *autonomy*, which leads to considerations of informed consent, privacy, and confidentiality, (2) *beneficence* and nonmaleficence, which involves considerations of risk/benefit and scientific merit, and (3) *justice*, which deals with such issues as compensation if there is injury in the trial or if there are benefits that the participants share in them. The key to justice is fairness—if a sense of unfairness is felt, there is reason to investigate whether justice has been withheld.

IRB approval is typically given for 1 year at a time. Researchers are required to get approval for protocol modifications, report protocol violations, and inform the committee of unanticipated side effects. The chair of an IRB, or staff person, may recommend "expedited review" for studies that meet certain minimal criteria.

Participants are guaranteed privacy of their data. Certain pieces of information such as birth date and location of birth are "protected information." Given such information, it could be possible to figure out the identity of the participant. Hence, there is a great deal of effort to *deidentify* the data. For example, the link to a participant's identity may be kept at a participating center with only a study number transmitted to a coordinating center.

A question is, who controls an IRB? While there are general guidelines, their implementation often depends on individual IRB members with a passion for a specific topic. The IRB review, like the local fire fighters' inventory of the premises, reflects what its members think important and it may be difficult to appeal a decision.

All researchers agree that ethical considerations take precedence over science. In practice, this may lead to valid differences of opinion. Also, new scientific procedures such as characterizing the human genome bring up new challenges and issues that require societal agreement as to what constitutes ethical behavior.

To reiterate what was stated at the beginning of this section, the principal investigator has the primary responsibility for the ethical conduct of a clinical trial. Institutional review boards, data safety monitoring boards, and committees internal to a particular study all assist with assuring ethical conduct.

7.7.3 Reporting

A look at journals such as the *Journal of the American Medical Association* indicates that there is a fairly standard approach to reporting the results of a clinical trial with enough information so that the validity and quality of the trial can be assessed. One of the more important characteristics of such reports is an accounting of all the subjects that had some role in the trial, starting with a pool of potential patients and ending with subjects enrolled and their progression through the trial. A useful tool for this purpose is the CONSORT diagram from the group, Consolidated Standards of Reporting Trials (http://www.consort-statement.org/). This diagram requires an explicit accounting of all the observations in a clinical trial. For a parallel group trial, Figure 7.2 lists the requirements. See Figure 7.3 for the report from the HSPP study.

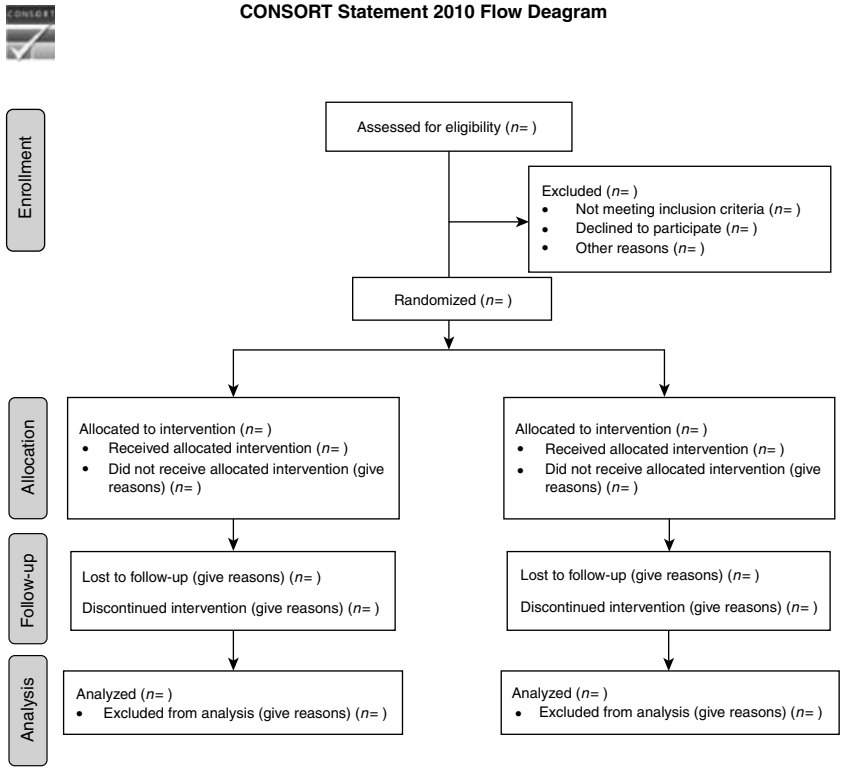

Figure 7.2 CONSORT diagram for accounting the disposition of subjects in a clinical trial (http://www.consort-statement.org/resources/downloads/).

The CONSORT group has also published a checklist for reporting of clinical trials (Figure 7.4). The checklist can be downloaded from the CONSORT website or found in many journals.

7.8 NOTES

7.8.1 Multicenter Trials

RCTs are frequently multicenter studies; given a small effect size and the large number of subjects required, one center cannot supply the required number of subjects in the time frame of the study. A second reason is robustness of results. Comparable outcomes among centers that vary in geography, patient composition, and idiosyncrasies of medical practice provide validity of the treatment. One drawback to multicenter studies is that the administrative effort increases exponentially. The larger the study, the more robust the treatment design needs to be. This leads to completely randomized,

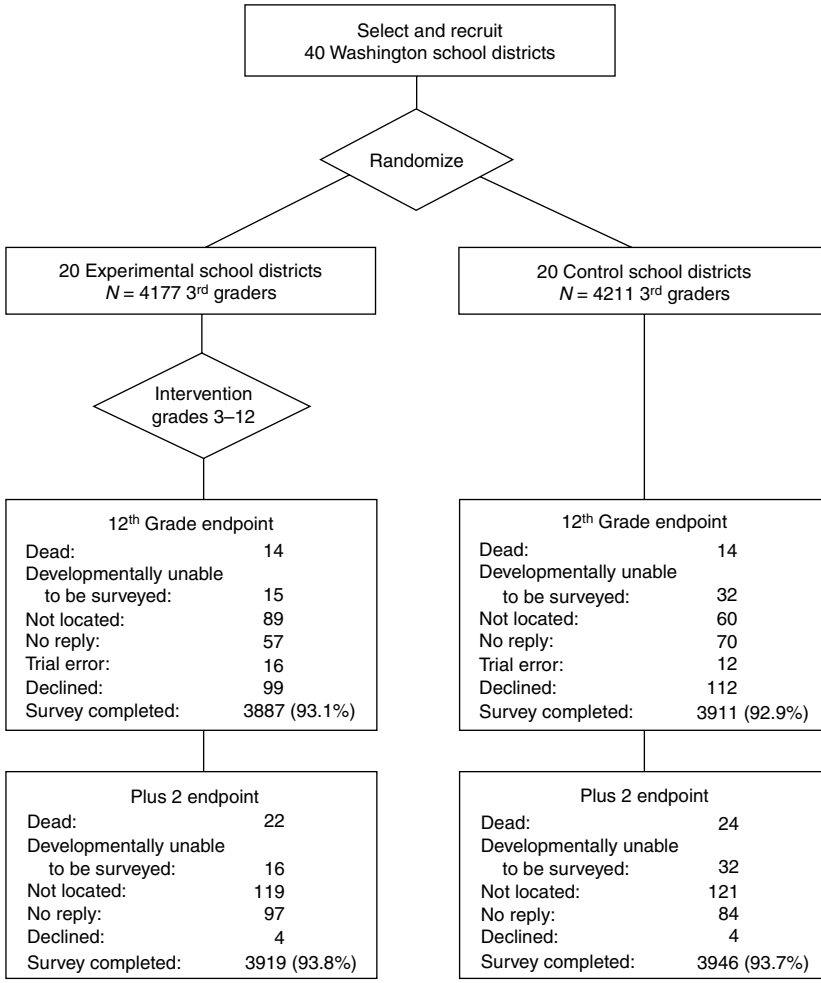

Figure 7.3 CONSORT diagram for HSPP study of Table 7.1. Copyright *Journal of the National Cancer Institute*, 2000. All rights reserved.

randomized block, or factorial designs. Latin square designs are rare. We are aware of fractional factorials or other designs common in industrial experimentation.

Multicenter trials usually have one or more coordinating centers. A coordinating center is the central nervous system of the clinical trial. It receives "messages" from the supervising groups such as the Steering Committee, stores data from the participating centers, does the multitude of tasks associated with data collection and processing, and sends out reports to these and other stakeholders. The first task of a coordinating center is to ensure standardization, specification and definition of data to be collected, and collection processes. This standardization requires a huge amount of time, travel, and training. There is a great deal of time pressure on center staff between the start

CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	
Participants	4a	Eligibility criteria for participants	
	4b	Settings and locations where the data were collected	
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	
	6b	Any changes to trial outcomes after the trial commenced, with reasons	
Sample size	7a	How sample size was determined	
	7b	When applicable, explanation of any interim analyses and stopping guidelines	
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	
	11b	If relevant, description of the similarity of interventions	

Figure 7.4 CONSORT checklist.

Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	
	13b	For each group, losses and exclusions after randomisation, together with reasons	
Recruitment	14a	Dates defining the periods of recruitment and follow-up	
	14b	Why the trial ended or was stopped	
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	
Other information			
Registration	23	Registration number and name of trial registry	
Protocol	24	Where the full trial protocol can be accessed, if available	
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	

*We strongly recommend reading this statement in conjunction with the CONSORT 2010 Explanation and Elaboration for important clarifications on all the items. If relevant, we also recommend reading CONSORT extensions for cluster randomised trials, non-inferiority and equivalence trials, non-pharmacological treatments, herbal interventions, and pragmatic trials. Additional extensions are forthcoming: for those and for up to date references relevant to this checklist, see www.consort-statement.org.

Figure 7.4 (*continued*)

of funding of the trial and enrollment of participants. There is a continuous tug of war between the desire to "improve" the trial by better definition of variables or collection of new variables (perhaps based on new scientific evidence) and the need for maintaining the original protocol. This requires not only knowledge on the part of center personnel but also wisdom.

7.8.2 International Harmonization

Given the international character of the pharmaceutical industry, it is clearly advantageous to harmonize and coordinate the development of new pharmaceutical. This effort is spearheaded by ICH:

> The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) is unique in bringing together the regulatory authorities and pharmaceutical industry of Europe, Japan and the US to discuss scientific and technical aspects of drug registration. Since its inception in 1990, ICH has evolved, through its ICH Global Cooperation Group, to respond to the increasingly global face of drug development, so that the benefits of international harmonisation for better global health can be realized worldwide. ICH's mission is to achieve greater harmonisation to ensure that safe, effective, and high quality medicines are developed and registered in the most resource-efficient manner.
>
> (http://www.ich.org/)

One of the key products of this effort is the publication of Efficacy Guidelines that are concerned with the design, conduct, safety, and reporting of clinical trials. The guidelines can be found at the ICH website. They are numbered E1 to E16 (so far). From a statistical point of view, the most important are ICH E3 (1995), ICH E9 (1999), and ICH E10 (2000). ICH E9 (1999) discusses criteria for establishing equivalence, among other topics. Reading and understanding these three guidelines will give you a very good sense of issues in clinical trials.

7.8.3 Data Safety Monitoring

Given the length of clinical trials, it is important that there be careful monitoring during data collection. The Women's Health Initiative is a good example of the benefits of monitoring with detection of mortality that was not expected but provided for. The unit that deals with these issues is a *Data Safety Monitoring Board* (DSMB). The principal functions of these boards are to monitor efficacy, safety, and approve modifications to the study protocol. Boards meet at least once a year, given written reports that are shared with IRBs.

One task of the DSMB is to carry out prespecified interim analyses. As indicated in the discussion of the Casa Pia study, an interim analysis uses up a small amount of the Type I error. A small amount of a Type I error ensures that it is very unlikely to reject the null hypothesis if it is true. By the same token, a very small Type I error

usually means a large Type II error and little power against a specified alternative. But it protects against very large deviations of the alternative from the null.

The DSMB can have access to the treatment assignment in the case of a study using blinding—as is the case in most trials. Some researchers have argued that even the DSMB should be blinded unless there is an emergency.

If a long-duration clinical trial continues year after year, there is of course information: no spectacular results can be expected.

7.8.4 Ancillary Studies

One form of data sharing is through *ancillary studies*. An ancillary study takes advantage of the basic structure of the clinical trial and adds another component. Colloquially, they are known as piggyback studies. A good example is the SELECT trial dealing with men over the age of 50. It turns out that selenium and vitamin E are also potential agents for preventing Alzheimer's disease. Hence, it was natural to consider adding measures of cognitive functioning to the SELECT trial. This led to the ancillary study, PREADVISE. Participants in the SELECT trial were invited to join that study as well. This involved additional informed consent, agreement by center directors to take part, and the establishment of a whole new data collection scheme.

Like an environmental impact investigation, ancillary study proposers have to justify that their study will not pose an undue burden on participants. The original investigators will be zealous to maintain the integrity of their study. This issue and others are carefully considered by IRBs—which are of course involved from the start.

7.8.5 Subgroup Analysis and Data Mining

Given the huge expense and volume of data coming from clinical trials, there is a logical impetus to "mine" these data. A distinction is often made between *primary* and *secondary* analyses. Primary analyses report the key results dealing with the reason for conducting the study. Secondary analyses may involve subgroups, secondary endpoints, or examination of the data suggested perhaps by new basic science findings. Sponsors of large clinical trials—most commonly a government agency—are keen to make "full use" of these data and grants for secondary analyses are now routinely available.

There are several challenges—and partial solutions—to subgroup analysis and data mining. First, a classic rule of statistical inference is that the same data should not be used for hypothesis generation and hypothesis testing. This makes sense. Second, there are approaches that allow some flexibility, for example, an analysis suggested independently in the literature, or split the data in two and use the first part for exploration and the second part for confirmation. Finally, the context of subgroup analyses should be part of a report of the results. A very nice graphical display of the analysis of 27 subgroups, specified beforehand, can be found in Howard et al. (2006). The use and misuse of subgroup analyses is discussed in Assmann et al. (2000). A good statistical reference is Berry (1990).

7.8.6 Meta-Analysis

Given that there are large number of trials that may deal with the same treatment or endpoint, there is a need for a methodology to combine the information. One approach is that of *meta-analysis*. Such analyses take into account the precision of each study and combine the results with, it is hoped, increased precision. There are many challenges to valid inference in such analyses beginning with subject selection, standardization of treatments, standardization of data collection, and standardization of endpoint measurements. The paradigm for a valid meta-analysis is a carefully conducted multicenter randomized clinical trial. Deviations from the paradigm are threats to valid inference. This is a huge area of current activity with publications in almost every issue of a medical journal. For an interesting example, see a meta-analysis of meta-analyses in Roseman et al. (2011). A statistical discussion can be found in Sutton and Higgins (2008) who discuss the "art and science" of meta-analysis.

7.8.7 Authorship and Recognition

The majority of public health RCTs are directed and guided by members of the academy where publication is the coin of the realm. Publication guidelines and principles are a crucial part of these trials. There usually is a key paper, which has been labeled the *initial trial publication*, that summarizes the results of the study. The first author usually is the principal investigator of the study followed by an entourage of coauthors with a footnote in fine print of all the principal investigators at the participating sites with their colleagues. The initial trial publication on diet and the risk of invasive breast cancer by WHI (Prentice et al., 2006) lists 47 coauthors in the masthead. Given the many years each of these researchers spent on the study, this is appropriate but modest recognition for all their work—but of little reward for a starting assistant professor. In these types of publications, the senior investigator, if not the writer of the paper, is usually listed last!

During the course of the trial, papers can be written about methodological aspects, characteristics of the participants, review of current status of the research area with particular reference to the trial, and—more rarely—reports on modification of the trial. Such publications are usually reviewed by a publication committee that makes sure that there are no references to current status of the endpoints.

7.8.8 Communication

A clinical trial involves a multitude of stakeholders: subjects, investigators, clinical and coordinating centers, review committees, advisory committees, sponsors (government, industry), and finally, the news media (which always wants to be first). Each entity has its own priorities, deadlines, and objectives. This requires careful prospective attention to content, context, and timing of communication—especially when unexpected findings turn up.

7.8.9 Data Sharing

It is agreed by all that data sharing is good, collegial, and scientifically useful. In practice, there are many obstacles. The three largest ones are the ongoing nature of research, privacy issues, and concerns about misuse of the data. A clinical trial typically produces a key results paper and data are not shared until after the publication.

Given the long time to carry out the trial, there are many secondary papers. In addition, there are many participating investigators who expect to have access to the data for their own research program. This leads to reluctance to sharing.

Some privacy issues have been discussed already. A uniquely contemporary issue is genetic information—which may constitute the ultimate identifier in criminal investigations! Sharing of genetic information presents new challenges.

While the design of a trial is straightforward, data collection and storage are not. The first issue is that standardization across centers is a complex activity. Detailed definitions of variables, their values, and exceptions take up volumes. The database has a complicated relational structure. The request for a "flat file" may be received with some scorn by the data managers. This leads to concern that the data requester may not really understand the intricacies of the data and may draw inappropriate conclusions.

Disposition and archiving of the data is now considered part of an application for a grant for a clinical trial. Prentice et al. (2005) give the conditions under which a "limited access database" from the WHI will be shared. First, a 3-year period between initial trial publication and sharing of the data. Second, a local Institutional Review Board has to approve the request. Third, manuscripts resulting from the analysis of the shared data need to be submitted to the sponsor of the WHI (National Heart Lung and Blood Institute of the United States) for review and comment prior to publication.

All this sounds rather daunting. And it is. However, given goodwill and some altruism, it is possible to share and most researchers are quite willing to do so. However, requests of "just send me the whole data set" will not be received kindly.

7.8.10 *N*-of-1 Trials

In contrast to the elaborate and expensive multicenter, multipatient, and multiinvestigator trials are the *N-of-1 trials.* An *N*-of-1 trial is simply a study on a single patient. Such studies have an honorable and distinguished history. Fisher (1971) begins the discussion of the design of experiments with a tea-tasting lady who can discriminate, she claims, between two ways of preparing tea. Fisher designs a study to assess the validity of her claim. In the health sciences, such studies may be appropriate in patients with chronic conditions. Larson et al. (1993) describe a series of such studies in patients with conditions such as chronic cough, atopic dermatitis, Parkinson's disease, and chronic headache. In a typical trial, active treatment would be compared with a placebo in a blinded fashion. Each trial had four to six sessions in each arm (replicates), each session lasting from 1 day to 4 weeks. Of 34 completed trials, 17 gave definitive results. Recently, the CONSORT group has become interested in such trials and will be reporting on such trials in 2011 (Vohra et al., 2011).

7.9 RESOURCES

As indicated, the field of clinical trials has exploded in the past 50 years. Among societies are the *Society for Clinical Trials* (http://www.sctweb.org/).

Journals devoted to clinical trial methodology include *Contemporary Clinical Trials* (formerly known as *Controlled Clinical Trials*), *Statistics in Medicine*, *Biometrics*, and *Journal of the American Statistical Association*.

Statistical texts focusing on clinical trials—in order of statistical depth—include Piantadosi (2005), Friedman et al. (2010), and Cook and DeMets (2008).

As mentioned, in the United States all clinical trials must be registered with the government (ClinicalTrials.gov). Leading medical journals will not publish papers from clinical trials unless the trial was registered before it was started. A primary reason is to prevent publication bias.

To get a broad overview of the field of clinical trials, review of the ICH E series documents is very useful. As mentioned, particularly, ICH E3 (1995), ICH E9 (1999), and ICH E10 (2000).

7.10 SUMMARY

The principles of public health and pharmaceutical clinical trial methodology are well established at this point in time. Some key principles are

- Adherence to government guidelines for the protection of human subjects. In the United States, the Office for Human Research Protection (www.hhs.gov/ohrp).
- Review and approval by an Institutional Review Board.
- A control group to be compared with active therapy.
- Randomization to ensure fair and unbiased comparison groups.
- Blinding to avoid introducing bias.
- Endpoints that have relatively permanent clinical relevance.
- Adequate planning for follow-up at the design stage—including funding.
- Adequate power.
- Per-protocol analyses of primary and secondary endpoints.
- Sponsors willing to fund the enterprise.
- Registration with a government agency before the start of the trial.

7.11 PROBLEMS

1. RCTs have been criticized from a variety of viewpoints (see Problem II in Chapter 1). Here is a comment by Bellomo and Bagshaw (2006) in the journal *Critical Care*, "Randomized trials, especially if associated with complex and strict protocols and many exclusion criteria, often give us the ability to know much but only

about a world that does not exist. Large observational studies, on the other hand, carry much uncertainty about causality but do describe the 'real' world. Likewise, observational studies have the distinct advantage of examining the long-term effects or prognosis of an intervention and assessing for adverse or rare outcome events."

(a) Comment on this quote. List other advantages and disadvantages of clinical trials and observational studies.

(b) A clinical trial must demonstrate both efficacy and safety. It has been said that clinical trials are good for determining efficacy but inadequate for demonstrating safety—observational studies are superior to clinical trials. Do you agree or disagree? Give specific reasons.

(c) Dr. Gordon Pledger, a former researcher with the U.S. Food and Drug Administration, has said that clinical trials do not reflect clinical practice—i.e. *effectiveness* rather than *efficacy*. Is this a reasonable summary of the Bellomo and Bagshaw quote? So why do clinical trials at all?

2. Grove (2011)—the former CEO of Intel—in an editorial in *Science* describes the clinical trial system in the United States as "Byzantine" and disappointing in output. Grove proposes that the FDA be only responsible for Phase I trials that emphasize safety. After that, the marketplace would take over with patient responses stored in huge databases that are now feasible. These databases could be accessed very quickly and the response of any patient or group of patients could be tracked in the database. He writes, "this would liberate drugs from the tyranny of the averages that characterize trial information today."

(a) If possible, access the editorial in *Science*.

(b) Given what you have learned in this chapter about clinical trials, comment on some scientific challenges: especially randomization, who gets into the database and how, maintenance of the database, quality of the data, incorporation of longitudinal data (since many treatments now involve chronic diseases), subgroup analyses, standardization of endpoints, and reporting of adverse events.

(c) Given that there are lots of negatives that could be said, state some positive aspects of this proposal.

3. (After Julious (2004)) A two-group RCT is planned to see whether a new drug is better at reducing blood pressure than a standard, well-established, drug. A reduction of 8 mmHg is considered meaningful. The standard deviation in the population of interest is about 40 mmHg.

(a) Assuming equal group size and a power of 0.80 and Type I error, 0.05, calculate the sample size per group needed for the study.

(b) A clinical trial begins with recruitment of subjects who may or may not consent to take part in the study. Suppose it is estimated that 75% of potential subjects will agree to participate in the study. In the scenario of part (a), how many subjects will have to be contacted?

(c) There is also the problem of dropouts during the trial. Suppose the dropout rate in part (a) is estimated to be 15%. How many subjects are needed for the trial? How many subjects will have to be recruited?

(d) The investigator wants to be sure that the study will pick up this clinically meaningful difference and wants the power to be 0.90. Recalculate the sample size and compare with your previous answer.

(e) The precision of the study is expressed by square of the standard error of the difference in the means (S.E.^{2}). In general, this will be (assuming equal allocation)

$$\text{S.E.}^2 = \sigma^2 \left(\frac{1}{n} + \frac{1}{n} \right) = \sigma^2 \left(\frac{2}{n} \right). \tag{7.1}$$

Suppose we now want to allocate different sample sizes to the two treatments, say, $n_T = rn_S$, where n_T is the number of subjects in the test treatment and n_S is the number of subjects in the standard treatment. To have equal precision in the two studies, we need to have

$$\frac{2}{n} = \frac{1}{n_S} + \frac{1}{n_T} = \frac{1}{n_S} + \frac{1}{rn_S}. \tag{7.2}$$

Solve this equation for n_S and show that

$$n_S = \frac{n}{2} \left(1 + \frac{1}{r} \right). \tag{7.3}$$

Finally, show that the total sample size for the study, instead of $2n$, is now

$$\text{Total sample size} = 2n \left[\frac{1}{4} \left(2 + r + \frac{1}{r} \right) \right]. \tag{7.4}$$

(f) Using equation 7.4, make a graph of total sample on the y-axis and r on the x-axis. Describe the behavior of the graph.

(g) Since the effect of the standard treatment is known, it is decided to put more effort into examining the new treatment and a decision is made to enroll twice as many subjects in the new treatment, keeping the same precision. Assuming the scenario in part (a), how many subjects need to be recruited for the new and standard treatments? What has happened to the total number of subjects to be recruited?

(h) One of the considerations in conducting a clinical trial is cost. Suppose that a clinical trial is conducted to compare two drugs, S and T. Suppose the cost of drug S is c_S and the cost of drug T is c_T. The question is how should sample sizes be allocated? It can be shown that the sample sizes should be allocated

via the *square root* rule (see, for example, van Belle (2008)),

$$\frac{n_S}{n_T} = \sqrt{\frac{c_T}{c_S}}. \tag{7.5}$$

The rule says to allocate sample sizes in the inverse of the square root of the ratio of the costs. Call this ratio r_c so that $n_S = n_T r_c$. This allows us to use the above equations to calculate sample sizes. In the situation of part (a), assume that the cost of the drugs for treatment S is \$40 and the cost for drugs in treatment T is \$640. How should sample sizes be allocated? What is the total cost for drugs in the study under equal allocation? How does that compare with the cost under unequal allocation?

4. Stiell et al. (2008) describe a study from the Resuscitation Outcome Consortium (ROC) for the treatment of out-of-hospital cardiac arrest (OHCA). The usual treatment is cardiopulmonary resuscitation (CPR) involving compression of the chest to at least 5 cm at a rate of 100/min. Two strategies are to be compared: first—the standard treatment—do CPR for 20–60 s, then analyze the heart rhythm, and, if necessary, shock the heart with a defibrillator (Analyze Early (AE)). Second, do CPR for 180 s before analyzing or shocking (Analyze Later (AL)). The endpoint of the study is a modified ranking score (MRS) of 3 or less at hospital discharge (labeled *neurologically intact*). The MRS ranges from 0 = no symptoms to 6 = dead. A score of 3 represents moderate residual disability. The proportion with MRS ≤ 3 is estimated to be 0.0541 for AE and 0.0745 with AL.

(a) Based on the observed proportions, calculate a "back-of-the-envelope" sample for a power of 0.80 and two-sided Type I error of 0.05, using the average of the two proportions for estimating the binomial variance (i.e., variance is estimated by $\bar{p}(1 - \bar{p})$).

(b) The clinical trial identifier is NCT00394706. Go to the registration website (http://www.clinicaltrials.gov) and locate this study. Note that the study is part of a larger study. Explore the history of this study by clicking the icon under More Information. Write a short paragraph summarizing the history of this study.

(c) As indicated, this study was part of a larger study. The effective sample size for this part of the study was 13,239. Assuming the same treatment effects and Type I error, calculate the actual power of the study.

(d) There was *cluster randomization* for this part of the study as follows. It was impractical to switch randomly between the two treatments: an EMS truck (rig) would do between 5 and 10 subjects in one arm of the study and then switch to the other arm. It was estimated that this would reduce the effective sample size by about 5%. Recalculate the power.

(e) Switching between the two arms was supervised by the ROC Coordinating Center. A few times, a site did not inform the center in time that the required number of subjects had been achieved and therefore continued enrolling in the current arm. The center statistician decided that subjects recruited after the

switch date would be counted in the other arm, based on the Intent to Treat strategy. The site investigators objected that this was unscientific. What is your opinion?

5. This problem deals with the two approaches to clinical trials discussed in Section 7.7.1: statistical significance and clinical importance. The clinical approach of the trial envisages a region where two treatments are considered equivalent.

(a) Prove that the clinical approach puts a more stringent requirement on proving that one treatment is better—rather than noninferior. That is, the sample sizes required are larger than those for the statistical approach.

(b) Go to ClinicalTrials.gov, select "search" and type in the word "noninferiority." How many studies are listed? Take the first 10 studies listed and determine who is the sponsor. What do you conclude?

(c) Now type in the word "equivalence." How many studies are listed? Why do you suppose this number is not the same as the number for noninferiority?

6. There is close link between the Type I error and the Type II error. In the pharmaceutical trial, the drug maker wants to maximize the power of the study while the regulatory agency wants to maintain the Type I error. The Type I error is called the *regulator risk*, and the Type II error is called the *producer risk* or *sponsor risk*.

(a) Interpret these errors in terms of approving or nor approving a new treatment.

(b) Why does the regulatory agency insist on maintaining a Type I error rate?

(c) Suggest at least two ways in which the drug maker can "fiddle" with the Type I error and, hence, increase the power of the study.

(d) Consider a study with one interim analysis in which the interim analysis is carried out at an α level of 0.01 and the final analysis is carried out at a level of 0.04. Prove that the overall α rate of the study is 0.05.

7. The mainstream medical journals will report at least one randomized clinical trial in a specific issue. Select a recent issue of the *Journal of the American Medical Association*, *Lancet*, *New England Journal of Medicine*, *British Medical Journal*, *Journal of the Canadian Medical Association*, or some other prestigious journal, and select a report of an RCT.

(a) Describe the experimental design under the headings of randomization, hypotheses, effect size, primary endpoint(s), and analyses.

(b) How were sample sizes determined?

(c) How restrictive were the eligibility criteria?

(d) If this was a multicenter study, how was the randomization carried out?

8. Subgroup analyses and data mining share the challenge of dealing with multiple looks at the same data set. Another common procedure is stepwise regression. Is this procedure subject to concerns about multiple testing? Why or why not?

9. The Hutchinson Smoking Prevention Project has been classified among "spectacular failures" (Patton et al., 2006). Given the remarkable feat of keeping track of 94% of 8388 enrolled third grade students for more than 9 years and strict adherence to the protocol, why should this study not be called a spectacular success? The issue, of course, is when a study shows a negative result whether that constitutes a failure or just shows that scientific research is not completely predictable. The HSPP authors argued (convincingly) that this study was very important in ruling out the reigning paradigm at the time of the start of the study that classroom social influences determine initiation and continuation of smoking.

(a) Discuss what constitutes a successful clinical trial.

(b) The Casa Pia study also demonstrated no statistically significant difference in treatments. Does this result differ from that of the HSPP?

(c) The SELECT study was terminated early because it was unlikely to show a significant treatment effect. Does such a study represent a *failure* or a *success* in view of the fact that the results were known earlier than expected with the opportunity to start a succeeding study earlier? Does the terminology of "stopping early" reflect a lack of equipoise about the treatments?

10. It has been argued that a physician can only consent to have a patient under their care take part in a randomized clinical trial if the physician is at *equipoise* about the treatments, that is, considers all the treatments of the trial equally effective (or, perhaps, ineffective). It has then been argued that no one is ever at complete equipoise and, hence, a physician can never refer a patient to a clinical trial involving randomization but must recommend the treatment he or she considers most likely to be effective.

(a) Discuss the validity of this argument.

(b) Fisher (1996) makes a distinction between emotional equipoise and scientific equipoise. In the above situation, emotional equipoise deals with the physician's personal feelings and preferences, for example, a reluctance to undergo general anesthesia. Fisher argues that this lack of emotional equipoise should not influence the scientific equipoise. Is this a valid distinction? Is it useful for the clinical trial recommendation?

11. As discussed in this chapter, using human subjects for experimentation requires safeguards. In this problem, a variety of scenarios are presented. How do the criteria of autonomy, beneficence, and justice enter in. If not, why not?

(a) It is not uncommon to pay human subjects for taking part in an experiment. Suppose a study requires 1 h of subject time. What would an IRB say about a payment of $20 for the subject's participation? What about a payment of $500?

(b) A teenager has agreed to take part in a phone interview, on depression. During the interview, the teenager expresses strong suicidal impulses. What should

the interviewer do in view of a guarantee of privacy to the teenager and the imminent threat of suicide? Which takes precedence? Why?

(c) Status epilepticus is a serious medical condition in which a subject arrives unconscious at an emergency department. Current treatment for the condition is primitive and unsatisfactory. Can the subject be assigned to an RCT investigating potentially beneficial therapies? If so, what safeguards must be in place?

(d) Many pharmaceutical clinical trials are *add-ons*, where a new treatment is added on to a therapy considered standard of practice. How can the new treatment be tried out as a *stand-alone*?

(e) A graduate student in industrial hygiene applies for human subjects approval to investigate two types of masks used by workers for controlling particulate emission during metal grinding. The Institutional Review Board is rather slow in reviewing the application but approves it. It learns subsequently that the student, under pressure of time, started the study before approval was given. What should the IRB do?

(f) Diesel exhaust contains known carcinogens. An occupational physician using an exposure chamber wants to expose volunteers to fairly substantial levels of standardized diesel exhaust in order to detect urinary biomarkers. The IRB refuses to give approval since the levels may be cancer inducing. The physician argues that the levels used are those found at downtown bus stops and that the exposure is shorter and carefully controlled. What should the IRB do?

(g) An undergraduate psychology student takes part in a study of "emotion." The purpose is to study frustration. Deception is used to create a sense of frustration. What happens to informed consent? Is deception ever permissible? Are there levels of deception? If deception is permissible and used, what are the obligations to the student at the end of the study?

(h) The Casa Pia school system in Lisbon was founded to serve orphans and homeless children 200 years ago.

i Currently, about 20% of the 4000 children are wards of the state and the director of the school system is their legal guardian. The issue was raised whether it was ethical for the director to have consent responsibility for these children enrolled in the study (about 100). Should this be an issue? If so, discuss and propose a solution.

ii The consent issue above was raised by the Data Safety Monitoring Board. Was this an appropriate concern of the DSMB given that two IRBs had reviewed and approved the study? In general, how should conflicting ethical judgments be resolved?

iii The study obtained informed consent from the parents. Technically, this is all that is needed for carrying out the study since the participants are minors. However, the investigators also obtained the *assent* of the children. What is the difference between assent and consent?

(i) The hormone replacement therapy arm of the Women's Health Initiative was terminated early due to excess mortality. A challenge was how to communicate the information to all the stakeholders: the approximately 30,000 study subjects, the physicians treating the patients, the principal investigators at each of the 40 participating centers, the company supplying the medication that ran the risk of lawsuits (as happened), and the news media. Discuss ethical and practical aspects.