CHAPTER 7

Clinical Trials

As indicated at the beginning of Chapter 2, some strange statements and strange ad hoc sample size determination formulas can be found in the literature. The medical literature is not immune to such problems, as Sathian, Sreedharan, Baboo, Sharan, Abhilash, and Rajesh (2010) used the words "proved" and "proven" in their article. In particular, they stated: "Thus, the study has to be repeated on a larger sample so that the real difference can be statistically proved." Nothing can ever be *proved* using statistics as there will always be an element of doubt since samples are taken from populations, rather than entire populations used (which would be highly impractical), so statements of that strength are improper. Such misstatements undoubtedly occur in many fields of application, however.

As the list of papers in the References at the end of this chapter indicates, a very large number of papers have been written on sample size determination for clinical trials. Some of these are tutorial-type papers that constitute recommended reading as supplements to this chapter and the books that have been written on the subject. For example, Thabane, Ma, Chu, Cheng, Ismalia, Rios, Robson, Thabane, Giangregorio, and Goldsmith (2010), Julious (2004), and Wittes (2002) are papers in this category. They can be useful supplements to the books that are listed in Section 7.1. Although not a tutorial paper on methodology, Bacchetti (2002) makes some important points about the refereeing of papers on sample size determination, suggesting that some research is not being published that should be published. Specifically, he stated: "Peer reviewers often make unfounded statistical criticisms, particularly in difficult areas such as sample size determination and power." These thoughts are echoed by Zwarenstein (2002) and others.

Lan and Shun (2009) discussed some important practical issues regarding clinical trials. In particular, they pointed out that the typical approach to sample size determination by solving for n such that there is the desired power of rejecting the

Sample Size Determination and Power, First Edition. Thomas P. Ryan.

null hypothesis when the true state of affairs is a specified alternative hypothesis is not necessarily what should be done. If two treatments are being compared, the null hypothesis is typically that their effects are equal; this being what the experimenter wants to reject. If that is done, however, the conclusion is that the new treatment is superior to the old treatment if the test is one-sided, or that the effects of the two treatments are unequal if the test is two-sided.

One interesting question is how much clinicians' beliefs should influence the sample size calculation. That is, should a trial committee include only people who are enthusiastic about a forthcoming trial or should the committee include a larger community of experts, including both enthusiasts and skeptics? This issue, which will likely influence sample size, was addressed by Fayers, Cushieri, Fielding, Uscinska, and Freedman (2000).

Simple superiority will generally be an inadequate conclusion, however, as the experimenter would like to conclude, with some high probability, that the true difference is at least equal to Δ_m, using the Lan and Shun (2009) terminology. Just setting the alternative hypothesis equal to Δ_m will not accomplish that, however. This problem can be remedied by simply setting the null and alternative hypotheses equal to H_0: $\Delta \leq \Delta_m$ and H_a: $\Delta > \Delta_m$, respectively, but the specification of such hypotheses seems to be uncommon in the medical literature, as well as in the applied literature, in general, although it is an established part of the theory of hypothesis testing.

In a relatively recent article that should be read by anyone involved in clinical trials and medical research, in general, as well as reading certain cited references, Bacchetti (2010) stated the following.

> Early in my career, an epidemiologist told me that dealing with sample size is the price one has to pay for being a biostatistician. Since then, I have spent untold time and effort paying this price, while also coming to realize that such effort produces no real scientific benefit. Unfortunately, widespread misconsceptions about sample size hurt not only statisticians, but also the quality of medical science generally .... I present here a wider challenge to current conventions, including how they cause serious harm .... Reports of completed studies should *not* [emphasis mine] include power calculations, and guidelines requiring them should be changed to instead discourage them.

In previous chapters, the difficulty of providing reasonable necessary input values was stressed, and this difficulty was especially apparent in Chapter 5 because of what had to be inputted. This is one of Bacchetti's main points, in addition to stating that the current approach "assumes a meaningful boundary between adequate and inadequate sample sizes that does not actually exist, not even approximately." He also pointed out that there is no indication of how a completed study's information should actually be used.

So what should be done instead? Bacchetti (2010) obviously feels that there should be a shift toward reporting point estimates and confidence intervals.

(Of course, many others feel the same way.) As he stated, "power is irrelevant for interpreting completed studies." Indeed, power is never really known because it depends on parameter values that are never known. This means, in turn, that sample sizes that are used are never known to be "correct."

Beyond that, Bacchetti (2010) stated: "A common pragmatic strategy is to use the maximum reasonably feasible . . . . When constraints imposed by funders determine feasibility, doing the maximum possible within those constraints is a sensible choice." The author also suggested considering the use of sample sizes that have worked well in the past for similar studies, pointing out that power calculations are frequently based on such studies anyway since preliminary data are frequently unavailable.

There are special considerations that must be made in determining sample size when the subjects are people, including the dropout rate and ethical considerations. These are discussed in Sections 7.1 and 7.2.

7.1 CLINICAL TRIALS

The determination of an appropriate sample size is especially important when the sample units are people. As discussed in Section 2.3, there are ethical considerations as it is important that a clinical study have sufficient power to detect an effect that would be deemed significant, as it would be unethical to expose the subjects to risks if a study was not likely to have a significant outcome.

There are various other factors that should be considered, including the dropout rate and time and resource constraints. Because of the considerable importance of determining an appropriate sample size when people are involved, more than a few books have been written on sample size determination for clinical trials, including Machin, Campbell, Tan, and Tan (2009), Julious (2009), Chow, Shao, and Wang (2008), and Lemeshow, Hosmer, Klar, and Lwanga (1990). See also Shuster (1992), Senn (2002), Kenward and Jones (2003) and the tables provided by Machin, Campbell, Fayers, and Pinol (1997), in addition to the material on sample size determination in Peace (2009).

The latter explained that although the FDA requires two clinical trials before approval of a new drug can be granted, under certain conditions a single clinical trial will suffice. Indeed, Darken and Ho (2004) argued for a single confirmatory trial and pointed out that use of a single trial has been gaining acceptance. If only one trial is used, it is obviously imperative that the trial be well designed, and of course the trial must show statistical significance. It is also highly desirable if only one trial is used for the p-value for the test of a new drug be extremely small, much less than what is needed to show statistical significance if $\alpha = .01$ or .05 is used. This is because of the reproducibility probabilities given in Table 1.12 of Chow et al. (2008, p. 6), which shows that there is approximately a 90% chance of a p-value less than .001 being observed in future clinical trials if the

p-value was .001 for the single clinical trial that was conducted. This contrasts with having only about a 50-50 chance of reproducing a p-value of .05 if that number was observed in the trial.

There are actually multiple sample sizes that need to be determined in clinical trials because there are multiple phases. Specifically, a Phase I study is a preliminary study that generally uses only a few patients (such as 5 or 6) at each dose level, with one of the objectives being to determine dosage level to be used in a later phase, such as the maximum tolerated dose levels; Phase II involves a moderate number of patients with an eye toward a preliminary assessment of new drug efficacy (without the use of a control group); and Phase III is intended to provide a definitive assessment of efficacy, often involving thousands of patients. There is usually a progression from Phase I to Phase II if only a specified minimum number of patients respond favorably to the drug. There can also be multiple stages within Phase II, with the outcome at each stage determining whether or not the next stage is reached. This is illustrated in Section 7.8.

As we would expect, Phase I and Phase II studies predominate because of the cost and effort involved in Phase III studies.

Parker and Berman (2003) described a clinical study in which one of them was involved that illustrates what often happens in the real world versus what happens in statistics books. In particular, they pointed out that the "variability known" assumption is frequently unfounded. Even if prior data exist, which will often not be the case, it may not be possible, or at least defensible, to use the estimate of σ from that data because that population may differ from the population under study, and different times and different conditions may also be important factors. The expected dropout rate must also be considered, and this may be difficult to estimate.

General considerations that should be made in determining sample size for clinical trials were given by Kirby, Gebski, and Keech (2002) and Eng (2003) is also an instructional article, written by an M.D., which mentions that many published clinical studies have suffered from low power. In an early paper, George and Desu (1974) considered the dual problem of determining sample size and the required duration of a trial for a fixed sample size. Makuch and Simon (1982) generalized those results.

An indirect way of determining sample size for clinical trials, by using simulation to determine the probability of observing a significant result for a given sample size and a given degree of superiority of the treatment relative to the control, was given by Allen and Seaman (2009). They approached this from a standpoint of both "noninferiority" and superiority, with the former defined as the absolute value of the control minus the new treatment being inside the confidence interval of $\pm k\%$, with k being the number that is judged to signify an important clinical difference. This simply means that the proposed treatment is not grossly inferior, if inferior at all, to the control. Of course, superiority would be signaled by the absolute difference of control minus treatment being greater than $k\%$, with

the outcome variable being such that a small percentage is desirable. Certainly simulations are only as good as the assumptions upon which they are based, but the charts and tables that the authors provided might be used as an aid in determining sample size. Eng (2004) also discussed the use of simulation to determine sample size, especially for complex designs. Similarly, Landau and Stahl (2013) also discussed the use of simulation for determining sample size, as they pointed out that complex modeling techniques might be used for which power formulas do not exist.

7.1.1 Cluster Randomized Trials

Cluster randomized trials, in which groups of individuals are randomized, are often used in the health field. Sample size determination must incorporate the effect of the clustering, which requires good estimates of intracluster correlation coefficients (ICCs). Campbell, Grimshaw, and Steen (2000) considered sample size determination for such studies and generated estimates of ICCs. You, Williams, Aban, Kabagambe, Tiwari, and Cutter (2011) considered sample size and power in the presence of variable cluster sizes.

7.1.2 Phase II Trials

Chang, Shuster, and Kepner (2004) considered Phase II clinical trials for which a binomial proportion was tested. They determined sample size for an exact unconditional test when the control group information has already been collected. Simon (1989) presented optimal two-stage designs for Phase II trials that were optimal in terms of minimizing the expected sample size for a given response probability, subject to certain constraints.

7.1.2.1 Phase II Cancer Trials

As explained by Chen (1997), the objective of a Phase II cancer clinical trial is to screen a treatment that can produce a similar or better response rate as can be achieved using standard methods. The screening is generally performed in two stages but Chen (1997) extended the procedure to three stages and considered sample size for each stage.

7.1.3 Phase III Trials

Gittins and Pezeshk (2002) discussed a "Behavioral Bayes" approach to sample size determination for Phase III clinical trials and for which a normal distribution is assumed. The optimal sample size is determined by minimizing the expected net cost as a function of the sample size. Chuang-Stein and Yang (2010) considered sample size decisions in the design of a Phase III superiority trial and argued that it is better to base the sample size decision for a confirmatory trial on the

probability that the trial will produce a positive outcome rather than use the traditional power approach. In an earlier, related paper, Chuang-Stein (2006) discussed the distinction between statistical power and the probability of having a successful trial and proposed an "average success probability." Richardson and Leurgans (1998) advocated a very reasonable approach for Phase III/IV clinical trials that clinical personnel and other professionals would do well to adopt. Specifically, they advocated doing a power analysis using a range of reasonable values or parameters that must be specified in order to determine sample sizes. Note that this is very similar to using the endpoints of a confidence interval for such parameters, as discussed in Section 2.1. De Martini (2010) considered the estimation of sample size for a Phase III trial based on Phase II data and De Martini (2011) considered the robustness of this approach when the effect size is lower in Phase III than in Phase II. Jiang (2011) considered optimal sample sizes for Phase II and Phase III work when decisions to move forward or not are based on probability of success.

7.1.4 Longitudinal Clinical Trials

Galbraith and Marschner (2002) considered longitudinal clinical trials and how to choose the sample size and frequency of measurement, with the objective being to minimize either the total number of measurements or the cost of the study. They proposed general design guidelines when there is dropout. See also Heo, Kim, Xue, and Kim (2010), who considered sample size for a longitudinal cluster randomized clinical trial.

7.1.5 Fixed Versus Adaptive Clinical Trials

In recent years there has been a push toward increased usage of adaptive clinical trials. That is, let the sample size be determined by results of a clinical trial as it progresses. Since sample size calculations can be a problem, one possibility would be to use adaptive trials rather than fixed trials. It is claimed that this reduces both the number of patients required and the length of time required for the trial. [See, for example, `http://pharmexec.findpharma.com/pharmexec/article/articleDetail.jsp?id=352793` and Jennison and Turnbull (2010).]

Adaptive designs are not without shortcomings, however. Chow and Chang (2008) provided a good review article of adaptive designs, presenting both the strengths and weaknesses and covering various types of adaptive designs. One of the weaknesses is that adaptive designs complicate matters somewhat and make the proper application of statistical methods more difficult. Another concern is that the actual patient population after the adaptations may differ from the original target population and the Type I error rate may not be controlled. Jahn-Eimermacher and Hommel (2007) indicated that an adaptive design won't always

be the best approach. Nevertheless, Tracy (2009, p. 117) stated that the use of adaptive designs is a "growing trend." Orloff, Douglas, Pinheiro, Levinson, Branson, Chaturvedi, Ette, Gatto, Hirsch, Mehta, Patel, Sabir, Springs, Stanski, Evers, Fleming, Singh, Tramontin, and Golub (2009) urged a move away from the traditional clinical development approach and toward an integrated approach that uses adaptive designs.

Software appropriately designed can facilitate the use of adaptive designs and make the use of such designs less difficult, thus overcoming the weakness stated by Chow and Chang (2008). Adaptive designs are in the same general spirit as pilot studies that were discussed in Section 2.1, which should be viewed positively, not negatively. Other papers on adaptive designs include Lehmacher and Wassmer (1999), Posch and Bauer (2000), Morgan (2003), Mehta and Patel (2006), Gao, Mehta, and Ware (2008), Coffey and Kairalla (2008), Bartroff and Lai (2008), Bretz, Koenig, Brannath, Glimm, and Posch (2009), and Lu, Chow, Tse, Chi, and Tang (2009), Mehta and Pocock (2011), and Wassmer (2011).

Sample size reestimation is possible as the clinical trial progresses, as discussed by Jennison and Turnbull (2003). There is a moderate amount of information on sample size reestimation in the literature, including review articles by Gould (2001), Chuang-Stein, Anderson, Gallo, and Collins (2006), and Proschan (2009). See also Gould (1995) and Herson and Wittes (1993).

7.1.6 Noninferiority Trials

Noninferiority trials are frequently conducted to justify the development of new drugs and vaccines. Chan (2002) developed a method for sample size and power calculations based on an exact unconditional test of noninferiority for testing the difference of two proportions. Friede and Stammer (2010) considered noninferiority clinical trials and presented a case study of a completely randomized active controlled trial in dermatology. Dann and Koch (2008) discussed methods that have been popular in the literature on noninferiority and discussed sample size considerations. Schwartz and Denne (2006) described a two-stage approach for sample size recalculation in noninferiority trials.

7.1.7 Repeated Measurements

Repeated measurements are often made in clinical trials. Sample size determination when there are repeated measurements has been addressed by Bloch (1986), Ahn, Overall, and Tonidandel (2001), Zhang and Ahn (2011), and Peters, Palmer, den Ruitjer, Grobbee, Crouse, O'Leary, Evans, Raichlen, and Bots (2012). Ahn and Jung (2005) investigated the implications of dropouts for the sample size estimates when a ramdomized parallel-groups repeated measurement design is used to compare two treatments. Lu, Luo, and Chen (2008) recognized subject attrition and proposed formulas for sample size estimation that take this into consideration.

Frison and Pocock (1992) considered, in particular, how to choose the number of baseline and post-treatment measurements. [Stata users may be interested in the command `sampsi2` for repeated measurement data, which implements the formulas in Frison and Pocock (1992).] See also Jung and Ahn (2003), who considered sample size determination with repeated measurements data when a generalized estimation equation approach is to be used for analysis. Dawson and Lagakos (1993) examined extensions of the approach of using a single summary statistic for each subject when there are repeated measures. Schouten (1999) also addressed the question of what statistics should be used when clinical trials have repeated measures.

7.1.8 Multiple Tests

There is often a need to perform multiple comparison tests with clinical trials data, such as when clinical trials are designed with multiple endpoints. Bang, Jung, and George (2005) presented a simple method for calculating sample size and power for a simulation-based multiple testing procedure that could be used in such situations. Senn and Bretz (2007) also considered sample size determination when there are multiple endpoints and multiple tests.

7.1.9 Use of Internal Pilot Studies for Clinical Trials

The general use of internal pilot studies as an aid in sample size determination was discussed in some detail in Section 2.1. The use of internal pilot studies in clinical trials work has been covered by Wittes and Brittain (1990), Birkett and Day (1994), and Proschan (2005), with the latter reviewing previous work. Kraemer, Mintz, Noda, Tinjlenberg, and Yesavage (2006) did urge caution in the use of pilot studies in clinical trials for guiding power calculations. Friede and Kieser (2003) discussed sample size reassessment in noninferiority and equivalence trials when an internal pilot study is used and Friede and Kieser (2011) considered an internal pilot study when an analysis of covariance is applied.

7.1.10 Using Historical Controls

Makuch and Simon (1980) developed a sample size formula for historical clinical trials, with the true control treatment effect considered to be equal to the observed effect from the historical control group. Many researchers subsequently pointed out, however, that the Makuch–Simon approach does not preserve the nominal power and Type I error due to the uncertainty in the true historical control treatment effect. This problem was addressed by Zhang, Cao, and Ahn (2010), who developed a sample size formula that properly accounts for the randomness in the observations from the historical control group. O'Malley, Norman, and

Kuntz (2002) used a Bayesian approach to determine the optimal sample size for a trial with a historical control.

7.1.11 Trials with Combination Treatments

Wolbers, Heemskerk, Chan, Yen, Caws, Farrar, and Day (2011) considered how best to separate the effects of combination treatments, as they compared a simple randomized trial of combination versus treatment with the use of a 2^2 factorial design. They concluded that, in the absence of interaction, an adequately powered 2^2 factorial design would require eight times as many observations as the combination trial.

7.1.12 Group Sequential Trials

A group sequential trial is a trial that allows for premature stopping due to safety, futility/efficacy, or both with options of additional adaptations based on results of an interim analysis. Jiang and Snapinn (2009) investigated the impact of nonproportional hazards on the power of group sequential methods. Chi, Hung, and Wang (1999) proposed a new group sequential test procedure that they claimed preserved the Type I error probability when there is sample size reestimation. Kim and DeMets (1992) considered sample size determination for group sequential trials when there is an immediate response and He, Shun, and Feng (2010) considered sample size when predefined group sequential stopping boundaries have to be adjusted. See also Mehta and Tsiatis (2001), Lai and Shih (2004), Jennison and Turnbull (2003, 2010), and, in particular, Chapter 8 of Chow et al. (2008).

7.1.13 Vaccine Efficacy Studies

Chan and Bohidar (1998) considered sample size determination for vaccine efficacy studies and compared the results obtained using two exact methods with the results obtained using a method based on a normal approximation.

7.2 BIOEQUIVALENCE STUDIES

Bassiakos and Katerelos (2006) proposed a sample size calculation for the case of therapeutic equivalence of two pharmaceuticals and Blackwelder (1993) addressed sample size and power for an equivalence trial of vaccines when there is interest in the relative risk of disease. Chow and Wang (2001) also considered sample size calculation for bioequivalence studies. They developed formulas for use with a crossover design and a parallel-group design with either raw data or log-transformed data. Siqueira, Whitehead, Todd, and Lucini (2005) compared

sample size formulas for a 2×2 crossover design applied to bioequivalence studies. See also Hauschke, Steinjans, Diletti, and Burke (1992).

7.3 ETHICAL CONSIDERATIONS

Ethical considerations were discussed generally in Section 2.3. This is especially important in clinical trials since people are involved, who should of course not be subjected to undue risks.

7.4 THE USE OF POWER IN CLINICAL STUDIES

Although Bacchetti (2010) decried the use of the word "power" in research articles describing the results of clinical trials, it is very unlikely that this practice will change anytime soon. It is often pointed out that trials involving human subjects could be considered unethical if the study is underpowered and there is some risk in participating in the clinical trial. This would especially be of concern with rapidly lethal diseases, with people participating in a trial for partly altruistic reasons. Horrobin (2003) discussed this scenario.

Bedard, Krzyzanowska, Pintillie, and Tannock (2007) surveyed all papers on two-arm Phase III randomized control trials presented at annual meetings of the American Society of Clinical Oncology during 1995–2003 and concluded that more than half of the randomized control trials with negative results did not have an adequate sample size to detect a medium-size treatment effect. [See also the discussion of this topic in Fayers and Machin (1995).]

Vickers (2003) investigated whether standard deviation values used in sample size calculations are smaller than those in the resulting study sample. The author concluded that there seems to be insufficient understanding that the standard deviation of a sample is a random variable and thus has variability and that standard deviations cannot easily be extrapolated from one population to another.

Obviously the results of Bedard et al. (2007) indicate a poor choice of sample size but Bacchetti (2010) takes issue with the position that such underpowered studies are unethical in stating: "The contention that inadequate power makes a study unethical . . . relies entirely on the threshold myth, a false belief that studies with less than 80% power cannot be expected to produce enough scientific or practical value to justify the burden imposed on participants. Because larger studies burden more participants, the fact of diminishing marginal returns implies that the ratio of projected value to total participant burden can only get worse with larger sample sizes. The risk of inadequate projected value relative to participant burden therefore applies to studies that are too large, not too small." [See Bacchetti, Wolf, Segal, and McCulloch (2005) for elaboration on this viewpoint.]

Bacchetti's position on power stands in stark contrast to the tenor of other articles, such as Whitley and Bell (2002), who were specifically concerned with the "hazard of under-powered studies," lamenting the fact that a study by Moher, Dulberg, and Wells (1994) revealed that for the results of 383 randomized controlled trials published in the *Journal of the American Medical Association, New England Journal of Medicine*, and *Lancet*, of the 102 null trials, only 16% had 80% power to detect a 25% relative difference between groups. Similarly, Freedman, Back, and Bernstein (2001) reviewed 717 manuscripts published in certain British and American orthopedic journals in 1997, from which 33 randomized, controlled trials were identified. Of the 25 studies which did not have an adequate sample size to detect a small effect (defined as 0.2 times the standard deviation), the average sample size used was only 10% of the required number. Thus, they concluded that inadequate sample sizes were being used in clinical orthopedic research.

Both positions have some merit. Clearly, power will always be unknown, both before and after a study has been conducted, and certainly some important information can still be gleaned from null studies. Not paying attention to power would certainly be unwise, however, as sample sizes should be large enough that there is a high probability of detecting an effect of such a magnitude that a scientist would want to detect.

7.5 PRECLINICAL EXPERIMENTATION

Tan, Fang, and Tian (2009) presented an experimental design for detecting departures from additivity of multiple drugs in preclinical studies and discussed sample size determination for their design.

7.6 PHARMACODYNAMIC, PHARMACOKINETIC, AND PHARMACOGENETIC EXPERIMENTS

Pharmacodynamic experiments are experiments involving test drugs that produce count measurements. Sample size determination for these and pharmacokinetic experiments have been addressed by Ogungbenro and Aarons (2010a,b) and Ogungbenro, Aarons, and Graham (2006). Pharmacogenetic experiments seek to identify the genetic factors that influence the intersubject variation in drug response. Tseng and Shao (2010) considered sample size determination for such studies. Sethuraman, Leonov, Squassante, Mitchell, and Hale (2007) gave sample size derivations for various designs that might be used in assessing or demonstrating pharmacokinetic dose proportionality.

7.7 METHOD OF COMPETING PROBABILITY

The method of competing probability, as introduced by Rahardja and Zhao (2009), was mentioned somewhat briefly in Section 2.2.4. The general idea is to determine sample size using this approach. For random variables X and Y chosen from the control and experimental treatments, respectively, the competing probability (*CP*) was defined by the authors as

$$\pi = \Pr(X < Y) + 0.5\Pr(X = Y)$$

Rahardja and Zhao (2009) stated that Bamber (1975) showed that the *CP* is equal to the area under the curve of the Receiver Operating Characteristics (ROC) curve, which is used in medical diagnostic testing.

Recall Eq. (3.6), which for a two-sided test would be

$$\begin{aligned} n &= \frac{(\sigma_1^2 + \sigma_2^2)(Z_{\alpha/2} + Z_\beta)^2}{(\Delta - \Delta_0)^2} \\ &= \frac{(2\sigma^2)(Z_{\alpha/2} + Z_\beta)^2}{(\Delta - \Delta_0)^2} \end{aligned} \tag{7.1}$$

under the assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and assuming that $n_1 = n_2 = n$, with Δ denoting $\mu_2 - \mu_1$ and Δ_0 denoting the value of $\mu_2 - \mu_1$ under the null hypothesis, which is typically zero.

Following Rahardja and Zhao (2009), the sample size can be written in terms of *CP* by first writing

$$\begin{aligned} \pi &= P(X < Y) + 0.5\ \Pr(X = Y) = \Pr(X - Y < 0) \\ &= P\left(\frac{X - Y - (-\Delta)}{\sigma\sqrt{2}} < \frac{\Delta}{\sigma\sqrt{2}}\right) \end{aligned}$$

with $\Delta/\sigma\sqrt{2}$ playing the role of $Z_{1-\pi}$ in terms of the general form of such probability statements Since the (standardized) effect size can be viewed as Δ/σ, it follows that $\Delta/\sigma = Z_{1-\pi}\sqrt{2}$. Letting Δ_0 in Eq. (7.1) $= 0$, the sample size formula in Eq. (7.1) can then be written as a function of $Z_{1-\pi}$:

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2}{Z_{1-\pi}^2} \tag{7.2}$$

Rahardja and Zhao (2009) stated that this formula would be used in conjunction with the question: "What is a clinically meaningful probability that the experimental treatment is better than the control?" The answer to that question would determine $Z_{1-\pi}$ and thus determine n.

To illustrate their proposed approach, the authors gave an example in which a clinical trial was to be used to compare the drug duloxetine with a placebo for the treatment of patients with diabetic peripheral neuropathic pain, with the measurement variable being the average weekly score of 24-hour average pain severity on an 11-point Likert scale. They referred to this study as described in Wernicke, Pritchett, D'Souza, Waninger, Tran, Iyengar, and Raskin (2006). Rahardja and Zhao (2009) stated that "it may be difficult to understand how an effect size of 0.545, as used by Wernicke et al. (2009), can be translated into the benefit that a patient may receive from duloxetine." With $\Delta/\sigma = Z_{1-\pi}\sqrt{2} = 0.545$, it follows that $Z_{1-\pi} = 0.385373$. Thus, $\Phi(0.385373) = \pi = .65$. With $\alpha = .05$ and a desired power of .90, the use of Eq. (7.2) produces

$$\begin{aligned} n &= \frac{(1.95996 + 1.28155)^2}{(0.385375)^2} \\ &= 70.75 \end{aligned}$$

so $n = 71$ would be used for each of the two groups.

7.8 BAYESIAN METHODS

As in other applications of sample size determination methods, there are Bayesian approaches that have been proposed for use in clinical trials, such as the methods proposed by Sahu and Smith (2006), Gajewski and Mayo (2006), Grouin, Coste, Bunouf, and Lecoutre (2007), Wang (2007), Patel and Ankolekar (2007), Brutti and De Santis (2008), Brutti, De Santis, and Gibbiotti (2008), Kikuchi, Pezeshk, and Gittins (2008), Whitehead, Valdés-Márquez, Johnson, and Graham (2008), Kikuchi and Gittins (2009), Cheng, Branscum, and Stamey (2010), and Zaslavsky (2009, 2012). Willan (2008) proposed a Bayesian approach that was designed to maximize expected profit, contingent upon the validity of the assumed model for expected total profit, and the method depending on the Central Limit Theorem. Two examples were given that illustrated the approach. Pezeshk (2003) gave a short review paper on Bayesian sample size determination techniques in clinical trials. See also Lan and Wittes (2012). Yin (2002) used a Bayesian approach for determining sample size for a proof of concept study, this being a study conducted by a pharmaceutical company for internal decision making.

Wang, Chow, and Chen (2005) proposed a Bayesian approach using a non-informative prior. It is obviously desirable to incorporate uncertainty into the parameter estimates that must be specified in sample size determination and from a practical standpoint an important question is how much difference this will make in the sample size. Wang et al. (2005) gave an example in which they compared the sample sizes obtained using their approach with the sample sizes

obtained using the standard frequentist formulas for testing equality, superiority, and equivalence. The standard formulas gave sample sizes of 273, 628, and 181, respectively, whereas their Bayesian approach produced corresponding sample sizes of 285, 654, and 188, respectively. Of course, the sample sizes using a Bayesian approach should be larger and here the percentage increases are 4.40, 4.14, and 3.87, respectively, so the percentage increases differ only slightly. Clearly, parameter estimate uncertainty should be incorporated in some manner and a Bayesian approach is one way to accomplish that.

7.9 COST AND OTHER SAMPLE SIZE DETERMINATION METHODS FOR CLINICAL TRIALS

Other nontraditional methods of sample size determination have been proposed for clinical trials. For example, Bacchetti, McCulloch, and Segal (2008) considered sample sizes based on cost efficiency. Bloch (1986) also considered cost, as did Briggs and Gray (1998), Kikuchi, Pezeshk, and Gittins (2008), Boyd, Briggs, Fenwick, Norrie, and Stock (2011), and Zhang and Ahn (2011). Cheng, Su, and Berry (2003) used decision analysis methods to arrive at sample sizes in each stage when a clinical trial is to be performed in two stages.

Another approach is the expected value of the information produced minus the total cost of the study. These are referred to as value of information (VOI) methods designed to determine the sample size so as to maximize the expected net gain. Such methods are described by Willan and Eckermann (2010). See also Willan and Pinto (2005) and Willan and Kowgier (2008).

7.10 META-ANALYSES OF CLINICAL TRIALS

Within the past 25 years, there has been much interest in meta-analysis, by which is meant combining data from individual studies and analyzing the data as if they had come from one huge study. This has been especially popular in clinical trials work, but the approach does have some limitations. DerSimonian and Laird (1986) examined eight published review articles, each of which reported results from several related trials. Clinical trials performed at different locations, under different conditions, and with differing numbers of subjects can't simply be merged in an unthinking manner. DerSimonian and Laird (1986) reported that Halvorsen (1986) found that only one article out of 589 examined considered combining results using formal statistical methods. DerSimonian and Laird (1986) proposed a random effects model for combining results from a series of experiments comparing two treatments. Jackson, Bowden, and Baker (2010) found, however, that it is inefficient for estimating the between-study variance

unless all of the studies that are combined are of similar size, but they did find that it is quite efficient for estimating the treatment effect. The DerSimonian and Laird model was extended by Jackson, White, and Thompson (2010). See also Shuster (2010) and the discussion of that paper.

7.11 MISCELLANEOUS

Internal pilot studies can certainly be useful but using results from other populations may not be helpful at all. Nevertheless, Yan and Su (2006) gave sample size formulas for a long-term trial involving patients with chronic disease by using results from existing short-term studies, as these may predict long-term disease progression patterns. Chen, DeMets, and Lan (2004) discussed the effect on the Type I error rate when the sample size is increased based on interim results and the study is unblinded. In contrast, Gould and Shih (1992) considered sample size reestimation *without* unblinding, but Waksman (2007) pointed out that some have claimed that the method can result in a severe underestimation of the within-group standard deviation.

Li, Shih, Xie, and Lu (2002) considered sample size adjustment in clinical trials based on conditional power and modified the procedure of Proschan and Hunsbarger (1995). Lachin (2005) reviewed methods for futility stopping based on conditional power. Lan, Hu, and Proschan (2009) discussed the use of conditional power and predictive power for early termination of a clinical trial. Spiegelhalter, Freedman, and Blackburn (1986) considered using interim results to obtain the predictive power of a trial, which would be obtained by averaging the conditional power (conditioned on the interim results) with the current belief about the unknown parameters. Dallow and Fina (2011), however, were critical of the use of predictive power and pointed out that the use of predictive power can lead to much larger sample sizes than either conditional power or standard sample size calculations. Wüst and Keiser (2003) considered possible adjustment of the sample size based on the value of the sample variance computed at an interim step. Xiong, Yu, Yan, and Zhang (2005) considered sample size for a trial for which the efficacy of a treatment is required for multiple primary endpoints. Liu and Dahlberg (1995) considered sample size requirements for K-sample trials with survival endpoints and $K = 3$ or 4. Wang, Chow, and Li (2002) derived sample size formulas for testing equality, noninferiority, superiority, and equivalence based on odds ratio for both parallel and crossover designs. Tango (2009) proposed a simple sample size formula for use with randomized controlled trials in which the endpoint is the count of recurrent events.

As a response to the question of whether a one-sided or two-sided test should be used, Dunnett and Gent (1996) proposed an alternative approach that tests simultaneously for a positive difference and for equivalence. Shen and Cai (2003)

considered sample size reestimation for clinical trials that involve censored survival data. Lin, Parks, Greshock, Wooster, and Lee (2011) focused on sample size calculations for clinical trial studies with a time-to-event endpoint in the presence of predictive biomarkers. Maki (2006) considered sample size determination when subjects are at risk for events other than the one of interest. Su (2005) determined sample size for endometrial safety studies that satisfies an FDA requirement and a requirement of the Committee on Proprietary Medicinal Products. Stalbovskaya, Hamadicharef, and Ifeachor (2007) presented an approach to sample size determination that involved estimation of probability density functions and confidence interval of parameters of a ROC (receiver operating characteristics) curve. Sozu, Sugimoto, and Hamasaki (2010, 2011, 2012) considered sample size determination when a trial has multiple co-primary endpoints. Fang, Tian, Li, and Tang (2009) considered sample size determination for trials with combinations of drugs. Sample size determination can be complicated by drug doses having different effects over populations in different regions. Zhang and Sethuraman (2010) considered this issue. Friede and Schmidli (2010a) developed a sample size reestimation strategy for count data in superiority and noninferiority trials that maintains the blinding of the trial. Friede and Schmidli (2010b) also considered sample size reestimation with count data with application to multiple sclerosis studies. Shun, He, Feng, and Roessner (2009) focused attention on sample size adjustment so as to maintain the Type I error rate. Shih (1993) and Shih and Zhao (1997) considered sample size reestimation for double-blinded trials with binary data. Shih (2009) devised a "perturbed unblinding" approach to sample size reestimation that keeps the treatment effect masked but allows an estimate of the variance using data from an interim stage. Hosmane, Locke, and Chiu (2010) considered power and sample size determination in QT/QTc studies.

McMahon, Proschan, Geller, Stone, and Sopko (1994) considered sample size determination relative to entry criteria for clinical trials that involve chronic diseases. Huang, Woolson, and O'Brien (2008) presented a sample size computation method for clinical trials with multiple outcomes and either O'Brien's (1984) test or its modification (Huang, Tilley, Woolson, and Lipsitz, 2005) is used for the primary analysis. Ivanova, Qaqish, and Schoenfeld (2011) gave a sample size formula for a sequential parallel comparison design. Kwong, Cheung, and Wen (2010) considered sample size determination when multiple comparisons of treatments are to be made with a control. Lakatos (1986) considered sample size determination when there are time-dependent rates of losses and noncompliance. Wu, Fisher, and DeMets (1980) considered sample size determination for long-term medical trials when there is time-dependent dropout.

Nomograms do not, in general, have the utility for sample size determination that they had 20 years ago because of current software capability. Software doesn't cover everything, however, so some researchers may find the nomogram of Malhotra and Indrayan (2010) useful, as it is for determining sample size for estimating sensitivity and specificity of medical tests.

7.12 SURVEY RESULTS OF PUBLISHED ARTICLES

The difficulties faced in trying to determine sample size have been discussed in previous chapters. This raises the question of how researchers actually arrive at sample size determinations, in clinical trials and in other types of studies.

Unfortunately, a recent study shows some serious problems and shortcomings. Specifically, Charles, Giraudeau, Dechartes, Bacon, and Ravaud (2009) "searched MEDLINE for all primary reports of two arm parallel group randomised controlled trials of superiority with a single primary outcome published in six high impact factor general medical journals between 1 January 2005 and 31 December 2006." The authors studied 215 articles, of which 5% did not report any sample size computation and 43% did not report all the parameters necessary for performing the computation.

Certain results from the study are somewhat distressing. For example, for the 157 reports that provided enough information for the sample size to be computed, the researchers found that their computed sample size and the reported sample size differed by more than 10% in 47 (30%) of the reports. Also disturbing was the fact that the difference between the assumptions for the control group and the observed data differed by more than 30% in 31% of the articles, and was greater than 50% in 17% of the articles. Only 34% of the articles reported all data that were necessary to compute the sample size, had an accurate sample size calculation, and used assumptions that were not refuted by the data. There was some guesswork done by the authors, however, as they assumed $\alpha = .05$ and a two-tailed test if the only missing information was the α-value and whether the test was one-sided or two-sided.

Of course, what would be most interesting would be an assessment of the proportion of studies with significant results that had these types of problems, and similarly for the proportion of studies that had nonsignificant results. For example, were nonsignificant results reported because errors resulted in a sample size being used that was too small (i.e., the study was underpowered)? This was not covered in the paper but it was covered in an online response by the first author. That author indicated that the results were similar for significant and nonsignificant results, with 55% of the studies with nonsignificant results reporting all the parameter values necessary for computing the sample size and this being done in 50% of the studies with significant results. The percentages with sample size computations that could be replicated were 80% and 73%, respectively. There is no indication of whether the authors were able to replicate the same numbers, however.

These results, although disturbing, should not be surprising in view of the difficulties noted in previous chapters. The numbers might actually be worse, however, as Charles et al. (2009) stated: "An important limitation of this study is that we could not directly assess whether assumptions had been manipulated to obtain feasible sample sizes because we used only published data." Such

(possible) manipulation was called "sample size samba" by Schulz and Grimes (2005). The need to report sample size calculations seems clear and has been urged by Sjögren and Hedström (2010).

7.13 SOFTWARE

The N Solution 2008 software (`http://www.pharmasoftware.net/products/n-solution`) is based on Chow, Shao, and Wang (2008). Java applets for Phase I and Phase II clinical trials are available at `http://biostats.upci.pitt.edu/biostats/ClinicalStudyDesign`. The latter is intended for use by the University of Pittsburgh Cancer Institute Biostatistics Facility only, but is also available for external users. See also the free software for clinical trials available at `http://www.cancerbiostats.onc.jhmi.edu/software.cfm` and the applet for a two-treatment crossover design or parallel design clinical trial at `http://hedwig.mgh.harvard.edu/sample_size/size.html`. (That applet can also be used to detect a relationship between a dependent variable and an independent variable, although the type of relationship to be detected is not clear.)

In addition to the software mentioned in this chapter, another prominent software package specifically for clinical trials is East by Cytel Software Corporation, which is apparently the most widely used clinical trial software system and has greater capabilities than other general-purpose commercial sample size determination software that is not just for clinical trials. The East-Adapt option permits midcourse sample size corrections, in the spirit of adaptive clinical trial designs. East 5 has considerable capabilities, as should be apparent from `www.cytel.com/pdfs/East_5_brochure_webFINAL.pdf`.

ExpDesign Studio is used for classical and adaptive clinical trial designs; its use is explained by Chang (2008).

Some of the popular sample size determination software mentioned repeatedly in previous chapters can also be used for clinical trials work. For example, PASS 11 has the capability for sample size determination for Phase II clinical trials: single-stage, two-stage, or three-stage.

For a single-stage Phase II clinical trial, PASS 11 determines sample size based on the work of A'Hern (2001), who provided tables for single-stage Phase II designs based on an exact test of proportions using the binomial distribution. To illustrate, the user would enter α and the desired power and enter the "maximum response rate of a poor treatment" (p_0, the null hypothesis) and the "minimum response rate of a good treatment" (p_1, the alternative hypothesis). For example, with $\alpha = .05$ and power $= .90$ for the one-sided test and $p_0 = .10$ and $p_1 = .20$, the software gives $n = 109$.

The algorithm that PASS 11 uses for two-stage Phase II clinical trials is based on and extends the work of Simon (1989). The algorithm will generally take

quite a while to run since it is a brute force search procedure. The output is far more extensive than the output for the single-stage case, as would be expected. To illustrate, again assume $p_0 = .10$, $p_1 = .20$, $\alpha = .05$, and power $= .90$. The detailed output is given below.

Two-Stage Clinical Trials Sample Size
Possible Designs For P0 = 0.100, P1 = 0.200, Alpha = 0.050, Beta = 0.100

N1	R1	PET	N	R	Ave N	Alpha	Beta	Constraints Satisfied
109	16	0.000	109	16	109.00	0.043	0.099	Single Stage
70	6	0.442	109	16	91.77	0.043	0.100	Minimax
42	4	0.588	121	17	74.56	0.049	0.099	Optimum

References

Simon, Richard. 'Optimal Two-Stage Designs for Phase II Clinical Trials',Controlled Clinical Trials, 1989, Volume 10, pages 1-10.

Report Definitions
N1 is the sample size in the first stage.
R1 is the drug rejection number in the first stage.
PET is the probability of early termination of the study.
N is the combined sample size of both stages.
R is the combined drug rejection number after both stages.
Ave N is the average sample size if this design is repeated many times.
Alpha is the probability of rejecting that P<=P0 when this is true.
Beta is the probability of rejecting that P>=P1 when this is true.
P0 is the response proportion of a poor drug.
P1 is the response proportion of a good drug.

Summary Statements
The optimal two-stage design to test the null hypothesis that P<=0.100 versus the alternative that P>=0.200 has an expected sample size of 74.56 and a probability of early termination of 0.588. If the drug is actually not effective, there is a 0.049 probability of concluding that it is (the target for this value was 0.050). If the drug is

```
actually effective, there is a 0.099 probability of
concluding that it is not (the target for this value was
0.100). After testing the drug on 42 patients in the first
stage, the trial will be terminated if 4 or fewer respond.
If the trial goes on to the second stage, a total of 121
patients will be studied. If the total number responding
is less than or equal to 17, the drug is rejected.
```

The symbols are defined in the section "Report Definitions." The program inputs should be made judiciously so that the output is not voluminous and the algorithm does not run for a long time—like for hours. (The latter can easily happen, as explained in the program's help file.)

For example, it will usually be desirable to use the "optimum designs" selection so that a large number of other designs that satisfy the constraints are not additionally printed out. When that selection is made, the output includes the single-stage design and the minimax design, in addition to the optimum design. The latter is the design that minimizes the expected sample size, $E(N)$, which is denoted in the output as "Ave N." This design is found through an exhaustive search of all possible designs. The minimax design is the design with the smallest total sample size. The reason there is an expected sample size rather than a known total sample size is the trial could terminate early after the first stage, and in this example that would happen if fewer than five patients responded to the drug. The expected sample size is computed, as would be expected, as $E(N) = N1 + (1 - PET)(N - N1)$. That is, it is the sample size in the first stage plus the second-stage sample size times the probability that the second stage is reached.

PASS 11 also has the capability for three-stage Phase II sample size determination, which is listed under "Proportions" in their menu. This is just an extension of a two-stage Phase II clinical trial, with a decision made at the end of the second stage as to whether or not to proceed to the third stage. Judicious choice of input selections is, of course, even more important for three stages than it is for two stages, so that the program doesn't run for hours and produce output that might be deemed unnecessary.

Stata 12 has the capability for sample size determination for clinical trials through various programs, including the menu-driven program of Royston and Babiker (2002), with this updated by Barthel, Royston, and Babiker (2005). See also Barthel, Royston, and Parmar (2009), who gave both menu and command-driven Stata programs for a time-to-event outcome with two or more experimental arms. A user-written command, `sampsi_fleming`, will determine sample size for a Fleming design (Fleming, 1982), which is a design for a Phase II clinical trial. For example, if the null hypothesis of $p_0 = 0.2$ is tested against an alternative of $p_1 = 0.4$ with $\alpha = .01$ and a target power of .90, the following output results when the command `sampsi_fleming, a(.01) p(.90)` is used.

```
Sample size calculation for the Fleming design
------------------------------------------------------
H0: p <= p0
H1: p > p0

p0 = .2 p1 = .4
With a sample size of 67
the null hypothesis is rejected if there are >= 22 respon-
ders
Type I error = 0.0093
Power        = 0.9082
```

Users of SAS Software may be interested in O'Brien and Castelloe (2007), which gives some illustrations of sample size analysis and power determination.

Although now outdated, Shih (1995) described the use of the software SIZE for determining sample size for clinical trials.

Software for sample size determination for clinical trials using Bayesian methods has not existed in the recent past and that is apparently still true.

7.14 SUMMARY

Although simple sample size formulas are desirable if hand calculations are necessary, "one size fits all" approaches should be avoided. Flaherty (2004) stated that 400 subjects is a reliable sample size for a study to have adequate statistical power. Fogel (2004) disputed this broad generalization, pointing out that of the first 40 abstracts of papers in the *Journal of the American Medical Association* that he looked at, 52% did not meet this requirement.

There is much to consider regarding sample size and power in clinical trials. Readers may be interested in the discussions in Proschan, Lan, and Wittes (2006), especially Chapter 3, which is entitled "Power: Conditional, Unconditional and Predictive," and Bacchetti (2010) is also recommended, as stated previously.

REFERENCES

A'Hern, R. P. (2001). Sample size tables for exact single-stage Phase II designs. *Statistics in Medicine*, **20**, 859–866.

Ahn, C. and S.-H. Jung (2005). Effect of dropouts on sample size estimates for test on trends across repeated measurements. *Journal of Biopharmaceutical Statistics*, **15**, 33–41.

Ahn, C., J. E. Overall, and S. Tonidandel (2001). Sample size and power calculations in repeated measurement analysis. *Computer Methods andPrograms in Biomedicine*, **64**(2), 121–124.

Allen, I. E. and C. A. Seaman (2009). Predicting success: Simulation can forecast probable success in clinical trials. *Quality Progress*, February, 60–63.

Bacchetti, P. (2002). Peer review of statistics in medical research: The other problem. *British Medical Journal*, **324**, 1271–1273.

Bacchetti, P. (2010). Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine*, **17** (electronic journal).

Bacchetti, P., C. E. McCulloch, and M. R. Segal (2008). Simple, defensible sample sizes based on cost efficiency. *Biometrics*, **64**, 577–585. Discussion: **64**, 586–594.

Bacchetti, P., L. E. Wolf, M. R. Segal, and C. E. McCulloch (2005). Ethics and sample size. *American Journal of Epidemiology*, **161**, 105–110.

Bamber, D. C. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic curve graph. *Journal of Mathematical Psychology*, **12**, 387–415.

Bang, H., S.-H. Jung, and S. George (2005). Sample size calculation for simulation-based multiple-testing procedures. *Journal of Biopharmaceutical Statistics*, **15**, 957–967.

Barthel, F.M.-S., P. Royston, and A. Babiker (2005). A menu-driven facility for complex sample-size calculation in randomized control trials with a survival or a binary outcome: Update. *The Stata Journal*, **5**(1), 123–129.

Barthel, F.M.-S., P. Royston, and M. K. B. Parmar (2009). A menu-driven facility for sample-size calculation in novel multiarm, multistage, randomized controlled trials with a time-to-event outcome. *The Stata Journal*, **9**(4), 505–523.

Bartroff, J. and T. L. Lai (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Statistics in Medicine*, **27**, 1593–1611.

Baskerville, N. B., W. Hogg, and J. Lemelin (2001). The effect of cluster randomization on sample size in prevention research. *Journal of Family Practice*, **50**, 241–246.

Bassiakos, Y. and P. Katerelos (2006). Sample size calculation for the therapeutic equivalence problem. *Communications in Statistics: Simulation and Computation*, **35**, 1019–1026.

Bedard, P. L., M. K. Krzyzanowska, M. Pintillie, and I. F. Tannock (2007). Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology Annual Meetings. *Journal of Clinical Oncology*, **25**(23), 3482–3487.

Birkett, M. A. and S. J. Day (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine*, **13**, 2455–2463.

Blackwelder, W. C. (1993). Sample size and power in prospective analysis of relative risk. *Statistics in Medicine*, **12**, 691–698.

Bloch, D. A. (1986). Sample size requirements and the cost of a randomized clinical trial with repeated measurements. *Statistics in Medicine*, **5**(6), 663–667.

Boyd, K. A., A. H. Briggs, E. Fenwick, J. Norrie, and S. Stock (2011). Power and sample size for cost-effectiveness analysis: fFN neonatal screening. *Contemporary Clinical Trials*, **32**(6), 893–901.

Bretz, F., F. Koenig, W. Brannath, E. Glimm, and M. Posch (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, **28**(8), 1181–1217.

Briggs, A. and A. Gray (1998). Power and sample size calculations for stochastic cost-effectiveness analysis. *Medical Decision Making*, **18**, Supplement, S81–S92.

Brutti, P. and F. De Santis (2008). Robust Bayesian sample size determination for avoiding the range of equivalence in clinical trials. *Journal of Statistical Planning and Inference*, **138**, 1577–1591.

Brutti, P., F. De Santis, and S. Gibbiotti (2008). Robust Bayesian sample size determination in clinical trials. *Statistics in Medicine*, **27**, 2290–2306.

Campbell, M., J. Grimshaw, and N. Steen (2000). Sample size calculations for cluster randomised trials. *Journal of Health Services Research Policy*, **15**, 12–16.

Chan, I. S. F. and N. Bohidar (1998). Exact power and sample size for vaccine efficacy studies. *Communications in Statistics—Theory and Methods*, **27**(6), 1305–1322.

Chan, I.-S. (2002). Power and sample size determination for noninferiority trials using an exact method. *Journal of Biopharmaceutical Statistics*, **12**(4), 457–469.

Chang, M. (2007). *Adaptive Design Theory and Implementation using SAS and R*. Boca Raton, FL. Chapman & Hall, CRC.

Chang, M. (2008). *Classical and Adaptive Clinical Trial Designs Using ExpDesign Studio*. Hoboken, NJ: Wiley.

Chang, M. N., J. J. Shuster, and J. L. Kepner (2004). Sample sizes based on exact unconditional tests for Phase II clinical trials with historical controls. *Journal of Biopharmaceutical Statistics*, **14**, 189–200.

Charles, P., B. Giraudeau, A. Dechartes, G. Bacon, and P. Ravaud (2009). Reporting of sample size calculations in randomised clinical trials: Review. *British Medical Journal*, **338**, 1732.

Chen, T. T. (1997). Optimal three-stage designs for Phase II cancer clinical trials. *Statistics in Medicine*, **16**, 2701–2711.

Chen, Y. H. J., D. L. DeMets, and K. K. G. Lan (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine*, **23**, 1023–1038.

Cheng, D., A. J. Branscum, and J. D. Stamey (2010). A Bayesian approach to sample size estimation for studies designed to evaluate continuous medical tests. *Computational Statistics and Data Analysis*, **54**(2), 298–307.

Cheng, Y., F. Su, and D. A. Berry (2003). Choosing sample size for a clinical trial using decision analysis. *Biometrika*, **90**, 923–926.

Chi, L., H. M. J. Hung, and S.-J. Wang (1999). Modification of sample size in group sequential trials. *Biometrics*, **55**, 853–857.

Chow, S. C. and M. Chang (2006). *Adaptive Design Methods in Clinical Trials.* New York: Chapman and Hall.

Chow, S.-C. and M. Chang (2008). Adaptive design methods in clinical trials—A review. *Orphanet Journal of Rare Diseases*, **3**(11), (electronic journal).

Chow, S.-C. and H. Wang (2001). On sample size calculation in bioequivalence trials. *Journal of Pharmacokinetics and Pharmacodynamics*, **28**, 155–169.

Chow, S.-C., J. Shao, and H. Wang (2008). *Sample Size Calculations in Clinical Research*, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

Chuang-Stein, C. (2006). Sample size and the probability of a successful trial. *Pharmaceutical Statistics*, **5**(4), 305–309.

Chuang-Stein, C. and R. Yang (2010). A revisit of sample size decisions in confirmatory trials. *Statistics in Biopharmaceutical Research*, **2**(2), 239–248.

Chuang-Stein, C., K. Anderson, P. Gallo, and S. Collins (2006). Sample size reestimation: A review and recommendations. *Drug Information Journal*, **40**, 475–484.

Coffey, C. S. and J. A. Kairalla (2008). Adaptive clinical trials: Progress and challenges. *Drugs in R & D*, **9**(4), 229–242.

Dallow, N. and P. Fina (2011). The perils with the use of predictive power. *Pharmaceutical Statistics*, **10**(4), 311–317.

Dann, R. S. and G. G. Koch (2008). Methods for one-sided testing of the difference between two proportions and sample size considerations related to non-inferiority clinical trials. *Pharmaceutical Statistics*, **7**, 130–141.

Darken, P. F. and S.-Y. Ho (2004). A note in sample size savings with the use of a single well-controlled clinical trial to support the efficacy of a new drug. *Pharmaceutical Statistics*, **3**, 61–63.

Dawson, J. D. and S. W. Lagakos (1993). Size and power of two-sample tests of repeated measures data. *Biometrics*, **49**(4), 1022–1032.

De Martini, D. (2010). Adapting by calibration the sample size of a Phase III trial on the basis of Phase II data. *Pharmaceutical Statistics*, **10**(2), 89–95.

De Martini, D. (2011). Robustness and corrections for sample size adaptation strategies based on effect size estimation. *Communications in Statistics—Simulation and Computation*, **40**(9), 1263–1277.

DerSimonian, R. and N. Laird (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, **7**, 177–188.

Dunnett, C. W. and M. Gent (1996). An alternative to the use of two-sided tests in clinical trials. *Statistics in Medicine*, **15**, 1729–1738.

Eng, J. (2003). Sample size estimation: How many individuals should be studied? *Radiology*, **227**, 309–313.

Eng, J. (2004). Sample size estimation: A glimpse beyond simple formulas. *Radiology*, **230**(3), 606–612.

Fang, H.-B., G.-L.Tian, W. Li, and M. Tang (2009). Design and sample size for evaluating combinations of drugs of linear and loglinear dose–response curves. *Journal of Biopharmaceutical Statistics*, **19**(4), 625–640.

Fayers, P. M. and D. Machin (1995). Sample size: How many patients are necessary? *British Journal of Cancer*, **72**, 1–9.

Fayers, K. W., A. Cushieri, J. Fielding, B. Uscinska, and L. S. Freedman (2000). Sample size calculation for clinical trials: The impact of clinician beliefs. *British Journal of Cancer*, **82**, 213–219.

Flaherty, R. (2004). A simple method for evaluating the clinical literature. *Family Practice Management*, **11**(5), 47–52 (May).

Fleming, T. R. (1982). One-sample multiple testing procedure for Phase II clinical trials. *Biometrics*, **38**, 143–151.

Fogel, J. (2004). Letter to the Editor. *Family Practice Management*, **11**(9), 14 (October).

Freedman, K. B., S. Back, and J. Bernstein (2001). Sample size and statistical power of randomised, controlled trials in orthopaedics. *The Journal of Bone and Joint Surgery*, **83-B**(3), 397–402.

Friede, T. and M. Kieser (2003). Blinded sample size reassessment with non-inferiority in non-inferiority and equivalence trials. *Statistics in Medicine*, **22**(6), 995–1007.

Friede, T. and M. Kieser (2011). Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical Statistics*, **10**(1), 8–13.

Friede, T. and H. Schmidli (2010a). Blinded sample size reestimation with negative binomial counts in superiority and non-inferiority trials. *Methods of Information in Medicine*, **49**(6), 618–624.

Friede, T. and H. Schmidli (2010b). Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis. *Statistics in Medicine*, **29**(10), 1145–1156.

Friede, T. and H. Stammer (2010). Blinded sample size recalculation in noninferiority trials: A case study in dermatology. *Drug Information Journal*, **44**(5), 599–607.

Frison, L. and S. J. Pocock (1992). Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design. *Statistics in Medicine*, **11**(13), 1685–1704.

Gajewski, B. J. and M. S. Mayo (2006). Bayesian sample size calculations in Phase II clinical trials using a mixture of informative priors. *Statistics in Medicine*, **25**, 2554–2566.

Galbraith, S. and I. C. Marschner (2002). Guidelines for the design of clinical trials with longitudinal outcomes. *Controlled Clinical Trials*, **23**(3), 257–273.

Gao, P., C. R. Mehta, and J. H. Ware (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics*, **18**, 1184–1196.

George, S. L. and M. M. Desu (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases*, **27**(1/2), 15–24.

Gittins, J. C. and H. Pezeshk (2002). A decision theoretic approach to sample size determination in clinical trials. *Journal of Biopharmaceutical Statistics*, **12**, 535–551.

Gould, A. L. (1995). Planning and revising the sample size for a trial. *Statistics in Medicine*, **14**, 1039–1051.

Gould, A. L. (2001). Sample size re-estimation: Recent developments and practical considerations. *Statistics in Medicine*, **20**, 2625–2643.

Gould, A. L. and W. Shih (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics—Theory and Methods*, **21**, 2833–2853.

Grouin, J.-M., M. Coste, P. Bunouf, and B. Lecoutre (2007). Bayesian sample size determination in non-sequential clinical trials: Statistical aspects and some regulatory considerations. *Statistics in Medicine*, **26**, 4914–4924.

Halvorsen, K.T. (1986). Combining results from independent investigations: Meta analysis in medical research. In *Medical Uses of Statistics* (J. C. Bailar III and F. Mosteller, eds.), pp. 392–418. Waltham, MA: NEJM Books.

Hauschke, D., V. W. Steinjans, E. Diletti, and M. Burke (1992). Sample size determination for bioequivalence assessment using a multiplicative model. *Journal of Pharmacokinetics and Biopharmaceutics*, **20**, 557–561.

He, Y., Z. Shun, and Y. Feng. (2010). Stopping boundaries of flexible sample size design with flexible trial monitoring. *Statistics in Biopharmaceutical Research*, **2**(3), 394–407.

Heo, M., Y. Kim, X. Xue, and M. Kim (2010). Sample size requirement to detect an intervention effect at the end of a follow-up in longitudinal cluster randomized trial. *Statistics in Medicine*, **29**, 382–390.

Herson, J. and J. Wittes (1993). The use of interim analysis for sample size adjustment. *Drug Information Journal*, **27**, 753–760.

Horrobin, D. F. (2003). Are large clinical trials in rapidly lethal diseases usually unethical? *Lancet*, **361**, 695–697.

Hosmane, B., C. Locke, and Y.-L. Chiu (2010). Sample size and power estimation in thorough QT/QTc studies with parallel group design. *Journal of Biopharmaceutical Statistics*, **20**(3), 595–603.

Huang, P., R. F. Woolson, and P. C. O'Brien (2008). A rank-based sample size method for multiple outcomes in clinical trials. *Statistics in Medicine*, **27**, 3084–3104.

Huang, P., B. C. Tilley, R. F. Woolson, and S. Lipsitz (2005). Adjusting O'Brien's test to control Type I error for the generalized nonparametric Behrens–Fisher problem. *Biometrics*, **61**(2), 532–538.

Ivanova, A., B. Qaqish, and D. A. Schoenfeld (2011). Optimality, sample size, and power calculations for the sequential parallel comparison design. *Statistics in Medicine*, **30**(23), 2793–2803.

Jackson, D., J. Bowden, and R. Baker (2010). How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*, **140**(4), 961–970.

Jackson, D., I. R. White, and S. G. Thompson (2010). Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine*, **29**, 1282–1297.

Jahn-Eimermacher, A. and G. Hommel (2007). Performance of adaptive sample size adjustment with respect to stopping criteria and time of interim analysis. *Statistics in Medicine*, **26**, 1450–1461.

Jennison, C. and B. W. Turnbull (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, **22**, 971–993.

Jennison, C. and B. W. Turnbull (2010). *Group Sequential and Adaptive Methods for Clinical Trials*, 2nd edition. New York: Chapman and Hall.

Jiang, K. (2011). Optimal sample sizes and go/no-go decisions for Phase II/III development programs based on probability of success. *Statistics in Biopharmaceutical Research*, **3**(3), 463–475.

Jiang, Q. and S. Snapinn (2009). Nonproportional hazards and the power of sequential trials. *Statistics in Biopharmaceutical Research*, **1**(1), 66–73.

Julious, S. A. (2004). Tutorial in biostatistics: Sample sizes for clinical trials with normal data. *Statistics in Medicine*, **23**, 1921–1986.

Julious, S. A. (2009). *Sample Sizes for Clinical Trials*. Boca Raton, FL: CRC Press.

Jung, S. H. and C. Ahn (2003). Sample size estimation for GEE method comparing slopes in repeated measurements data. *Statistics in Medicine*, **22**(8), 1305–1315.

Kenward, M. G. and B. Jones (2003). *Design and Analysis of Cross-Over Trials*. New York: Chapman and Hall.

Kikuchi, T. and J. Gittins (2009). A behavioral Bayes method to determine the sample size of a clinical trial considering efficacy and safety. *Statistics in Medicine*, **28**, 2307–2324.

Kikuchi, T., H. Pezeshk, and J. Gittins (2008). A Bayesian cost–benefit approach to the determination of sample size in clinical trials. *Statistics in Medicine*, **27**, 68–82.

Kim, K. and D. L. DeMets (1992). Sample size determination for group sequential clinical trials with immediate response. *Statistics in Medicine*, **11**, 1391–1399.

Kirby, A., V. Gebski, and A. C. Keech (2002). Determining the sample size in a clinical trial. *Medical Journal of Australia*, **177**(5), 256–257.

Kraemer, H. C., J. Mintz, A. Noda, J. Tinjlenberg, and J. A. Yesavage (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry*, **63**, 484–489.

Kwong, K. S., S. H. Cheung, and M.-J. Wen (2010). Sample size determination in step-up procedures for multiple comparisons with a control. *Statistics in Medicine*, **29**(26), 2743–2756.

Lachin, J. M. (2005). A review of methods for futility stopping based on conditional power. *Statistics in Medicine*, **24**(18), 2747–2764.

Lai, T. L. and M.-C. Shih (2004). Power, sample size and adaptation consideration in the design of group sequential clinical trials. *Biometrika*, **91**(3), 507–528.

Lakatos, E. (1986). Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clinical Trials*, **7**, 189–199.

Lan, K. K. G. and Z. Shun (2009). A short note on sample size estimation. S*tatistics in Biopharmaceutical Research*, **1**(4), 356–361.

Lan, K. K. G. and J. T. Wittes (2012). Some thoughts on sample size: A Bayesian-frequentist approach. *Clinical Trials*, **9**(5), 561–569.

Lan, K. K. G., P. Hu, and M. A. Proschan (2009). A conditional power approach to the evaluation of predictive power. *Statistics in Biopharmaceutical Research*, **1**(2), 131–136.

Landau, S. and D. Stahl (2013). Sample size and power calculations for medical studies by simulation when closed form expressions are not available. *Statistical Methods in Medical Research*, (to appear).

Lehmacher, W. and G. Wassmer (1999). Adaptive sample-size calculations in group sequential trials. *Biometrics*, **55**, 1286–1290.

Lemeshow, S., D. W. Hosmer, J. Klar, and S. K. Lwanga (1990). *Adequacy of Sample Size in Health Studies*. Chichester, UK: Wiley.

Li, G., W. J. Shih, T. Xie, and J. Lu (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics*, **3**(2), 277–287.

Lin, X., D. C. Parks, J. Greshock, R. Wooster, and K. R. Lee (2011). Effect of predictive performance of a biomarker for the sample size of targeted designs for randomized clinical trials. *Statistics in Biopharmaceutical Research*, **3**(4), 536–548.

Liu, P-Y. and S. Dahlberg (1995). Design and analysis of multiarm clinical trials with survival endpoints. *Controlled Clinical Trials*, **16**(2), 119–130.

Lu, K., X. Luo, and P.-Y. Chen (2008). Sample size estimation for repeated measures analysis in randomized clinical trials with missing data. *The International Journal of Biostatistics*, **4**(1), 1–16.

Lu, Q., S-. C. Chow, S. K. Tse, Y. Chi, and L. Y. Tang (2009). Sample size estimation based on event data for a two-stage survival adaptive trial with different durations. *Journal of Biopharmaceutical Statistics*, **19**, 311–323.

Machin, D., M. J. Campbell, S.-B. Tan, and S.-H. Tan (2009). *Sample Size Tables for Clinical Studies*. London : BMJ Books.

Machin, D., M. Campbell, P. Fayers, and A. Pinol (1997). *Sample Size Tables for Clinical Studies*. New York: Wiley.

Maki, E. (2006). Power and sample size considerations in clinical trials with competing risk endpoints. *Pharmaceutical Statistics*, **5**, 159–171.

Makuch, R. W. and R. M. Simon (1980). Sample size comparisons for non-randomised comparative studies. *Journal of Chronic Diseases*, **33**, 175–181.

Makuch, R. W. and R. M. Simon (1982). Sample size requirements for comparing time-to-failure among *k* treatment groups. *Journal of Chronic Diseases*, **35**(11), 861–867.

Malhotra, R. K. and A. Indrayan (2010). A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Research Methodology*, **58**(6), 519–522.

McMahon, R. P., M. Proschan, N. L. Geller, P. H. Stone, and G. Sopko (1994). Sample size calculations for clinical trials in which entry criteria and outcomes are counts of events. *Statistics in Medicine*, **13**(8), 859–870.

Mehta, C. R. (2011). Sample size reestimation for confirmatory clinical trials. In *Designs for Clinical Trials: Prospectives on Current Issues* (D. Harrington, ed.). New York: Springer.

Mehta, C. R. and N. R. Patel (2006). Adaptive, group, sequential and decision theoretic approaches to sample size determination. *Statistics in Medicine*, **25**, 3250–3269.

Mehta, C. R. and S. J. Pocock (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*, **30**(28), 3267–3284.

Mehta, C. R. and A. A. Tsiatis (2001). Flexible sample size considerations using information based interim monitoring. *Drug Information Journal*, **35**, 1095–1112.

Moher, D., C. S. Dulberg, and G. A. Wells (1994). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association*, **272**(2), 122–124. (Available at `http://www.ncbi.nlm.nih.gov/pubmed/8015121`.)

Morgan, C. C. (2003). Sample size re-estimation in group sequential response-adaptive clinical trials. *Statistics in Medicine*, **22**, 3843–3857.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079–1087.

O'Brien, R. G. and J. M. Castelloe (2007). Sample size analysis for traditional hypothesis testing: Concepts and issues. In *Pharmaceutical Statistics Using SAS: A Practical Guide* (A. Dmitrienko, C. Chuang-Stein, and R. D'Agostino, eds.), pp. 237–271, Chapter 10. Cary, NC: SAS.

Ogungbenro, K. and L. Aarons (2010a). Sample size/power calculations for population pharmacodynamic experiments involving repeated-count measurements. *Journal of Biopharmaceutical Statistics*, **20**(5), 1026–1042.

Ogungbenro, K. and L. Aarons (2010b). Sample-size calculations for multi-group comparison in population experiments. *Pharmaceutical Statistics*, **9**, 255–268.

Ogungbenro, K., L. Aarons, and G. Graham (2006). Sample size calculations based on generalized estimating equations for population pharmacokinetic experiments. *Journal of Biopharmaceutical Statistics*, **16**, 135–150.

O'Malley, A. J., S.-L. T. Norman, and R. E. Kuntz (2002). Sample size calculation for a historically controlled clinical trial with adjustment for covariates. *Journal of Biopharmaceutical Statistics*, **12**(2), 227–247.

Orloff, J., F. Douglas, J. Pinheiro, S. Levinson, M. Branson, P. Chaturvedi, E. Ette, P. Gatto, G. Hirsch, C. Mehta, N. Patel, S. Sabir, S. Springs, D. Stanski, M. R. Evers, E. Fleming, N. Singh, T. Tramontin, and H. Golub (2009). The future of drug development: Advancing clinical design. *Nature Reviews Drug Discovery*, **8**, 949–957.

Parker, R. A. and N. G. Berman (2003). Sample size: More than calculations. *The American Statistician*, **57**(3), 166–170.

Patel, N. R. and S. Ankolekar (2007). A Bayesian approach for incorporating economic factors in sample size design for clinical trials of individual drugs and portfolios of drugs. *Statistics in Medicine*, **26**, 4976–4988.

Peace, K. E. (2009). *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*. Boca Raton, FL: Chapman and Hall, CRC.

Peters, S. A., M. K. Palmer, H. M. den Ruitjer, D. E. Grobbee, J. R. Crouse III, D. H. O'Leary, G. W. Evans, J. S. Raichlen, and M. L. Bots (2012). Sample size requirements in trials using repeated measurements and the impact of trial design. *Current Medical Research and Opinion*, **28**(5), 681–688.

Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: A short review. *Statistical Methods in Medical Research*, **12**, 489–504.

Posch, M. and P. Bauer (2000). Interim analysis and sample size reassessment. *Biometrics*, **56**, 1170–1176.

Proschan, M. A. (2005). Two-stage sample size re-estimation based on a nuisance parameter—A review. *Journal of Biopharmaceutical Statistics*, **15**, 559–574.

Proschan, M.A. (2009). Sample size re-estimation in clinical trials. *Biometrical Journal*, **51**, 348–357.

Proschan, M. A. and S. A. Hunsbarger (1995). Designed extension of studies based on conditional power. *Biometrics*, **51**, 1315–1324.

Proschan, M., K. K. G. Lan, and J. T. Wittes (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. New York: Springer.

Rahardja, D. and Y. D. Zhao (2009). Unified sample size computations using the competing probability. *Statistics in Biopharmaceutical Research*, **1**(3), 323–327.

Richardson, D. J. and S. Leurgans (1998). Sample size justification in Phase III/IV clinical trials. *Neuroepidemiology*, **17**(2), 63–66.

Royston, P. and A. Babiker (2002). A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome. *The Stata Journal*, **2**(2), 151–163.

Royston, P. and F. M.-S. Barthel (2005). Projection of power and events in clinical trials with a time-to-event outcome. *The Stata Journal*, **10**(3), 386–394.

Sahu, S. K. and T. M. F. Smith (2006). A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **169**, 235–253.

Sathian, B., J. Sreedharan, N. S. Baboo, K. Sharan, E. S. Abhilash, and E. Rajesh (2010). Relevance of sample size determination in medical research. *Nepal Journal of Epidemiology*, **1**(1), 4–10.

Schouten, H. J. A. (1999). Planning group sizes in clinical trials with a continuous outcome and repeated measures. *Statistics in Medicine*, **18**(3), 255–264.

Schulz, K. F. and D. A. Grimes (2005). Sample size calculations in randomised trials: Mandatory and mystical. *The Lancet*, **365**, 1348–1353.

Schwartz, T. A. and J. S. Denne (2006). A two-stage sample size recalculation procedure for placebo- and active-controlled non-inferiority trials. *Statistics in Medicine*, **25**(19), 3396–3406.

Senn, S. J. (2002). *Cross-over Trials in Clinical Research*, 2nd edition. New York: Wiley.

Senn, S. and F. Bretz (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*, **6**, 161–170.

Sethuraman, V. S., S. Leonov, L. Squassante, T. R. Mitchell, and M. D. Hale (2007). Sample size calculation for the power model for dose proportionality studies. *Pharmaceutical Statistics*, **6**, 35–41.

Shen, Y. and J. Cai (2003). Sample size reestimation for clinical trials with censored survival data. *Journal of the American Statistical Association*, **98**(462), 418–426.

Shih, J. H. (1995). Sample size calculation for complex clinical trials with survival endpoints. *Controlled Clinical Trials*, **16**, 395–407.

Shih, W. J. (1993). Sample size re-estimation for triple blind clinical trials. *Drug Information Journal*, **27**, 761–764.

Shih, W. J. (2009). Two-stage sample size reassessment using perturbed unblinding. *Statistics in Biopharmaceutical Research*, **1**(1), 74–80.

Shih, W. J. and P. L. Zhao (1997). Design for sample size re-estimation with interim data for double blind clinical trials with binary outcomes. *Statistics in Medicine*, **16**, 1913–1923.

Shun, Z., Y. He, Y. Feng, and M. Roessner (2009). A unified approach to flexible sample size design with realistic constraints. *Statistics in Biopharmaceutical Research*, **1**(4), 388–398.

Shuster, J. J. (1992). *Practical Handbook of Sample Size Guidelines for Clinical Trials.* Boca Raton, FL: CRC Press.

Shuster, J. J. (2010). Empirical vs. natural weighting in random effects meta-analysis. *Statistics in Medicine*, **29**, 1259–1265. Discussion: **29**, 1266–1281.

Simon, R. (1989). Optimal two-stage designs for Phase II clinical trials. *Controlled Clinical Trials*, **10**, 1–10.

Siqueira, A. L., A. Whitehead, S. Todd, and M. M. Lucini (2005). Comparison of sample size formulae for 2×2 cross-over designs applied to bioequivalence studies. *Pharmaceutical Statistics*, **4**(4), 233–243. Discussion: **5**, 231–233.

Sjögren, P. and L. Hedström (2010). Sample size determination and statistical power in randomized controlled trials. Letter to the Editor. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology*, **109**(5), 652–653.

Sozu, T., T. Sugimoto, and T. Hamasaki (2010). Sample size determination in clinical trials with multiple co-primary binary endpoints. *Statistics in Medicine*, **29**(21), 2169–2179.

Sozu, T., T. Sugimoto, and T. Hamasaki (2011). Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *Journal of Biopharmaceutical Statistics*, **21**, 650–668.

Sozu, T., T. Sugimoto, and T. Hamasaki (2012). Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biometrical Journal*, **54**(5), 716–729.

Spiegelhalter, D. J., L. S. Freedman, and P. R. Blackburn (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, **7**, 8–17.

Stalbovskaya, V., B. Hamadicharef, and E. Ifeachor (2007). Sample size determination using ROC analysis. *Proceedings of the 3rd International Conference on Computational Intelligence in Medicine and Healthcare*, July 25–27. Plymouth, U.K.

Stone, G. W. and S. J. Pocock (2010). Randomized trials, statistics, and clinical inference. *Journal of American College of Cardiology*, **55**, 428–431.

Su, G. (2005). Sample size and power analysis for endometrial safety studies. *Journal of Biopharmaceutical Statistics*, **15**, 491–499.

Tan, M. T., H.-B. Fang, and G.-L. Tian (2009). Dose and sample size determination for multi-drug combination studies. *Statistics in Biopharmaceutical Research*, **1**(3), 301–316.

Tango, T. (2009). Sample size formula for randomized controlled trials with counts of recurrent events. *Statistics and Probability Letters*, **79**, 466–472.

Thabane, L., J. Ma, R. Chu, J. Cheng, A. Ismalia, L. P. Rios, R. Robson, M. Thabane, L. Giangregorio, and C. H. Goldsmith (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology*, **10**(1), open access journal.

Tracy, M. (2009). *Methods of Sample Size Calculations for Clinical Trials*. Master of Science thesis, University of Glasgow, Glasgow, Scotland. (Available at `http://theses.gla.ac.uk/671`.)

Tseng, C.-H. and Y. Shao (2010). Sample size analysis for pharmacogenetic studies. *Statistics in Biopharmaceutical Research*, **2**(3), 319–328.

Vickers, A. J. (2003). Underpowering in randomized trials reporting a sample size calculation. *Journal of Clinical Epidemiology*, **56**, 717–720.

Waksman, J. A. (2007). Assessment of the Gould–Shih procedure for sample size reestimation. *Pharmaceutical Statistics*, **6**(1), 53–65.

Wang, H. and S.-C. Chow (2007). Sample size calculation for comparing time-to-event data. In *Wiley Encyclopedia of Clinical Trials*, pp. 1–7. Hoboken, NJ: Wiley.

Wang, H., S.-C. Chow, and M. Chen (2005). A Bayesian approach on sample size calculation for comparing means. *Journal of Biopharmaceutical Statistics*, **15**, 799–807.

Wang, H., S.-C. Chow, and G. Li (2002). On sample size calculations based on odds ratios in clinical trials. *Journal of Biopharmaceutical Statistics*, **12**, 471–483.

Wang, H., B. Chen, and S.-C. Chow (2003). Sample size determination based on rank tests in clinical trials. *Journal of Biopharmaceutical Statistics*, **13**(4), 735–751.

Wang, M.-D. (2007). Sample size reestimation by Bayesian prediction. *Biometrical Journal*, **49**(3), 365–377.

Wassmer, G. (2011). On sample determination in multi-armed confirmatory adaptive designs. *Journal of Biopharmaceutical Statistics*, **21**(4), 802–817.

Wernicke, J. F., Y. L. Pritchett, D. N. D'Souza, A. Waninger, P. Tran, S. Iyengar, and J. Raskin (2006). A randomized controlled trial of duloxetine in diabetic peripheral neuropathic pain. *Neurology*, **67**, 1411–1420.

Whitehead, J., E. Valdés-Márquez, P. Johnson, and G. Graham (2008). Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*, **27**, 2307–2327.

Whitley, E. and J. Bell (2002). Statistics review: Sample size calculations. *Critical Care*, **6**(4), 335–341.

Willan, A. R. (2008). Optimal sample size determinations from an industry perspective based on the expected value of information. *Clinical Trials*, **5**(6), 587–594.

Willan, A. and S. Eckerman (2010). Optimal clinical trial design using value of information methods with imperfect implementation. *Health Economics*, **19**(5), 549–561.

Willan, A. R. and M. E. Kowgier (2008). Determining optimal sample sizes for multi-stage randomized clinical trials using value of information methods. *Clinical Trials*, **5**, 289–300.

Willan, A. R. and E. M. Pinto (2005). The expected value of information and optimal clinical trial design. *Statistics in Medicine*, **24**(12), 1791–1806. (Correction: **25**, 720.)

Wittes, J. (2002). Sample size calculations for randomized controlled trials. *Epidemiologic Reviews*, **24**(1), 39–53.

Wittes, J. and E. Brittain (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, **9**, 65–72.

Wolbers, M., D. Heemskerk, T. T. H. Chan, N. T. B. Yen, M. Caws, J. Farrar, and J. Day (2011). Sample size requirements for separating out the effects of combination treatments: Randomised controlled trials of combination therapy vs. standard treatment compared to factorial designs for patients with tuberculous meningitis. *Trials*, **12**, 26.

Wu, M., M. Fisher, and D. DeMets (1980). Sample sizes for long-term medical trials with time-dependent dropout and event rates. *Controlled Clinical Trials*, **9**, 119–136.

Wust, K. and M. Keiser (2003). Blinded sample size recalculation for normally distributed outcomes using long- and short-term data. *Biometrical Journal*, **45**(8), 915–930.

Xiong, C., K. Yu, Y. Yan, and Z. Zhang (2005). Power and sample size for clinical trials when efficacy is required in multiple endpoints: Application to an Alzheimer's treatment trial. *Clinical Trials*, **2**(5), 387–393.

Yan, X. and X. Su (2006). Sample size determination for clinical trials in patients with nonlinear disease progression. *Journal of Biopharmaceutical Statistics*, **16**, 91–105.

Yin, Y. (2002). Sample size calculation for a proof of concept study. *Journal of Biopharmaceutical Statistics*, **12**, 267–276.

You, Z., O. D. Williams, I. Aban, E. K. Kabagambe, H. K. Tiwari, and G. Cutter (2011). Relative efficiency and sample size for cluster randomized trials with variable cluster sizes. *Clinical Trials*, **8**(1), 27–36.

Zaslavsky, B. G. (2009). Bayes models of clinical trials with dichotomous outcomes and sample size determination. *Statistics in Biopharmaceutical Research*, **1**(2), 149–158.

Zaslavsky, B. G. (2012). Bayesian sample size estimates for one sample test in clinical trials with dichotomous and countable outcomes. *Statistics in Biopharmaceutical Research*, **4**(1), 76–85.

Zhang, S. and C. Ahn (2010). Effects of correlation and missing data on sample size estimation in longitudinal clinical trials. *Pharmaceutical Statistics*, **9**, 2–9.

Zhang, S. and C. Ahn (2011). Adding subjects or adding measurements in repeated measurement studies under financial constraints. *Statistics in Biopharmaceutical Research*, **3**(1), 54–64.

Zhang, S., J. Cao, and C. Ahn (2010). Calculating sample size in trials using historical controls. *Clinical Trials*, **7**, 343–353.

Zhang, W. and V. Sethuraman (2010). On power and sample size calculation in ethnic sensitivity studies. *Journal of Biopharmaceutical Statistics*, **21**(1), 18–23.

Zwarenstein (2002). Letter to the Editor and discussion. *British Medical Journal*, **325**, 491.

EXERCISES

7.1. Consider a single-stage Phase II clinical trial with desired power of .90 and $\alpha = .05$. The maximum response rate of a poor treatment is .10 and the minimum response rate of a good treatment is .25. Use PASS 11 or other software to determine the required sample size for the clinical trial. For what numbers of responses will the null hypothesis be rejected?

7.2. Why should there be concern over survey results of research articles that give the results of clinical trials?

7.3. What is one advantage that the method of competing probability has from the viewpoint of a patient over other methods for sample size determination that were discussed in the chapter?