Handbook of Statistics

VOLUME 27

General Editor

C.R. Rao

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Epidemiology and Medical Statistics

Edited by

C.R. Rao
Center for Multivariate Analysis
Department of Statistics, The Pennsylvania State University
University Park, PA, USA

J.P. Miller
Division of Biostatistics
School of Medicine, Washington University in St. Louis
St. Louis, MO, USA

D.C. Rao
Division of Biostatistics
School of Medicine, Washington University in St. Louis
St. Louis, MO, USA

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
North-Holland is an imprint of Elsevier

N·H

Handbook of Statistics, Vol. 27
ISSN: 0169-7161

DOI: 10.1016/S0169-7161(07)27016-6

16

Sequential and Group Sequential Designs in Clinical Trials: Guidelines for Practitioners

Madhu Mazumdar and Heejung Bang

Abstract

In a classical fixed sample design, the sample size is set in advance of collecting any data. The main design focus is choosing the sample size that allows the clinical trial to discriminate between the null hypothesis of no difference and the alternative hypothesis of a specified difference of scientific interest. A disadvantage of fixed sample design is that the same number of subjects will always be used regardless of whether the true treatment effect is extremely beneficial, marginal, or truly harmful relative to the control arm. Often, it is difficult to justify because of ethical concerns and/or economic reasons. Thus, specific early termination procedures have been developed to allow repeated statistical analyses to be performed on accumulating data and to stop the trial as soon as the information is sufficient to conclude. However, repeated analyses inflate the false positive error to an unacceptable level. To avoid this problem, many approaches of group sequential methods have been developed. Although there is an increase in the planned sample size under these designs, due to the sequential nature, substantial sample size reductions compared with the single-stage design is also possible not only in the case of clear efficacy but also in the case of complete lack of efficacy of the new treatment. This feature provides an advantage in utilization of patient resource. These approaches are methodologically complex but advancement in software packages had made the planning, monitoring, and analysis of comparative clinical trials according to these approaches quite simple. Despite this simplicity, the carrying on of a trial under group sequential design requires efficient logistics with dedicated team of data manager, study coordinator, biostatistician, and clinician. Good collaboration, rigorous monitoring, and guidance offered by an independent data safety monitoring committee are all indispensable pieces for its successful implementation.

In this chapter, we provide a review of sequential designs and discuss the underlying premise of all current methods. We present a recent example and an historical example to illustrate the methods discussed and to provide a flavor of the variety and complexity in decision making. A comprehensive list of

Newcombe, R.G. (1996). Sequentially balanced three-squares cross-over designs. *Statistics in Medicine* **15**, 2143–2147.

Neuhaus, J.M., Kalbfleisch, J.D., Hauck, W.W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25–35.

Patterson, H.D., Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545–554.

Patterson, S.D., Jones, B. (2006). *Bioequivalence and Statistics on Clinical Pharmacology*. Chapman & Hall/CRC, Boca Raton.

Prescott, P. (1999). Construction of sequentially counterbalanced designs formed from two latin squares. *Utilitas Mathematica* **55**, 135–152.

Russell, K.G. (1991). The construction of good change-over designs when there are fewer units than treatments. *Biometrika* **78**, 305–313.

Schuirmann, D.J. (1987). A comparison of the two one sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 657–680.

Senn, S.J. (1997). Cross-over trials. In: Ar-mitage, P., Colton, T. (Eds.),**vol. 2** *Encyclopedia in Biostatistics*. Wiley, New York.

Senn, S.J. (2000). Crossover design. In: Chow, S.C. (Ed.), *Encyclopedia of Biopharmaceutical Statistics*. Marcel Dekker, New York.

Senn, S.J. (2002). *Cross-over Trials in Clinical Research*, 2nd ed. Wiley, Chichester.

Senn, S.J., Lambrou, D. (1998). Robust and realistic approaches to carry-over. *Statistics in Medicine* **17**, 2849–2864.

Sheehe, P.R., Bross, I.D.J. (1961). Latin squares to balance immediate residual and other effects. *Biometrics* **17**, 405–414.

Stufken, J. (1996). Optimal Crossover Designs. In: Gosh, S., Rao, C.R. (Eds.), *Handbook of Statistics 13: Design and Analysis of Experiments*. Amsterdam, North-Holland, pp. 63–90.

Verbeke, G., Molenberghs, G. (2000). *Linear Mixed Models in Practice*. Springer, New York.

Williams, E.J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research* **2**, 149–168.

Zeger, S.L., Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous. *Biometrics* **42**, 121–130.

Zeger, S.L., Liang, K.Y., Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

softwares is provided for easy implementation along with practical guidelines. Few areas with potential for future research are also identified.

1. Introduction

Randomized clinical trial (RCT) is regarded as the gold standard for assessing the relative effectiveness/efficacy of an experimental intervention, as it minimizes selection bias and threats to validity by estimating average causal effects. There are two general approaches for designing RCT: (1) fixed sample design (FSD) and (2) group sequential design (GSD). In FSD, a predetermined number of patients (ensuring a particular power for proving a given hypothesis) are accrued, and the study outcome is assessed at the end of the trial. In contrast, a design where analyses are performed at regular intervals after a group of patients are accrued is called GSD. In comparative therapeutic trials with sequential patient entry, FSDs are often unjustified on ethical and economic grounds, and GSDs are preferred for their flexibility (Geller et al., 1987; Fleming and Watelet, 1989). Currently used methods can be classified into three categories: group sequential methods for repeated significance testing; stochastic curtailment or conditional power (Lan et al., 1982; Pepe and Anderson, 1992; Betensky, 1997) and Bayesian sequential methods (Spiegelhalter and Freedman, 1994; Fayers et al., 1997). While no single approach addresses all the issues, they do provide useful guidance in assessing the emerging trends for safety and benefit.

Trials using GSDs are common in published literature and the advantage of this kind of design is self evident by their impact (Gausche et al., 2000; Kelly et al., 2001; Sacco et al., 2001). One example of its successful use is a trial reported by Frustaci et al., where 190 sarcoma patients (a rare form of cancer) were to be accrued in order to detect a 20% difference in 2-year disease-free survival (60% on the adjuvant chemotherapy treatment arm versus 40% in the control arm undergoing observation alone) (Frustaci et al., 2001). An interim analysis was planned after half of the patients were accrued with stopping rule in terms of adjusted p-value. The trial was stopped as this criterion was met thereby saving 50% of the planned patient accrual. The observed difference was found to be 27% (72% on the treatment arm versus 45% on the control arm), 7% higher than what was hypothesized initially at the design stage. Therefore, the risk of treating additional patients with suboptimal therapy was greatly reduced.

Independent data safety monitoring committee (DSMC) with responsibilities of (1) safeguarding the interests of study patients, (2) preserving the integrity and credibility of the trial in order to ensure that future patients be treated optimally, and (3) ensuring that definitive and reliable results be available in a timely manner to the medical community has been mandated for all comparative therapeutic clinical trials sponsored by national institutes (URL: http://cancertrials.nci.nih.gov; Ellenberg, 2001). GSD provides an excellent aid to the DSMC for decision making. Other names utilized for this kind of committees playing virtually the same role are data or patient safety monitoring board (DSMB or PSMB), data monitoring and ethics committees (DMEC), and policy and data monitoring board (PDMB).

In this chapter, we start with a historical account of sequential methods and provide introduction to the underlying concept and approaches to the commonly utilized methods of inflation factor (IF) for sample size calculation and alpha spending function for monitoring the trials for early stopping. A listing of softwares is provided that has the capabilities of accommodating all of the methods discussed. A table of IF for sample size calculation of GSD is provided for quick assessment of feasibility of a trial (in regard to sample size) even before acquiring any special software for GSD. One current example is presented with standard template of a biostatistical consideration for writing study protocol, details of a stopping boundary utilized, items to be included in an interim analysis reports presented to the DSMC, and the substance included in the statistical section write-up for final dissemination in published literature. Another historical example (the BHAT trial) is discussed to highlight that the DSMC's decision to stop early was based not only on statistical group sequential boundary point, but also on a variety of other subjective considerations.

Several review papers and books from various perspectives are recommended to those who wish to learn about further details (Fleming and DeMets, 1993; Jennison and Turnbull, 2000; Sebille and Bellissant, 2003; Proschan et al., 2006).

2. Historical background of sequential procedures

The first strictly sequential method, the sequential probability ratio test, was developed during the Second World War (Wald, 1947). As its main application was the quality control of manufactured materials, its publication was only authorized after the end of the war, in 1947. Another class of sequential test is based on triangular continuation regions (Anderson, 1960). The basic idea on which these methods rely is to constantly use the available information to determine whether the data are compatible with null hypothesis, with alternative hypothesis, or insufficient to choose between these two hypotheses. In the first two cases, the trial is stopped and the conclusion is obtained whereas in the third case the trial continues. The trial is further processed until the data allows a legitimate (or per-protocol) decision between the two hypotheses. An example of a completely sequential trial can be found in Jones et al. (1982).

Armitage (1954) and Bross (1952) pioneered the concept of group sequential methods in medical field (Bross, 1952; Armitage, 1954). At first, these plans were fully sequential and did not gain widespread acceptance perhaps due to the inconvenience in their application. The problems discussed included the fact that response needs to be available soon after the treatment is started and that there would be organizational problems, such as coordination in multicenter trials and a much greater amount of work for the statistician. The shift to group sequential methods for clinical trials did not occur until the 1970s. Elfring and Schultz (1973) specifically used the term 'group sequential design' to describe their procedure for comparing two treatments with binary response (Elfring et al., 1973).

McPherson (1974) suggested that the repeated significance tests of Armitage et al. (1969) might be used to analyze clinical trial data at a small number of interim analyses (Armitage et al., 1969; McPherson, 1974). Canner (1977) used Monte Carlo simulation to find critical values of a test statistic for a study with periodic analyses of survival endpoint (Canner, 1977). However, Pocock (1977) was the first to provide clear guidelines for the implementation of the GSD attaining particular operating characteristics of type I error and power (Pocock, 1977). He made the case that most investigators do not want to evaluate results every time a couple of new patients are accrued but do want to understand the comparative merit every few months to assess if the trial is worth the time and effort and that continual monitoring does not have a remarkable benefit. More specifically, only a minor improvement is expected with more than five interim looks. A more comprehensive account of this history can be read from the excellent book by Jennison and Turnbull (2000).

3. Group sequential procedures for randomized trials

A primary difficulty in performing repeated analyses over time is the confusion about the proper interpretation of strength of evidence obtained from such evaluations. Suppose that only a single data analysis is performed after data collection has been fully completed for a trial. Then a two-sided (or one-sided if justified, e.g., non-inferiority design) significance value of $p \leq 0.05$, obtained from a test of hypothesis of no difference between an experimental therapy and a control, is usually interpreted as providing strong enough evidence that the new therapy provides an advantage. The interpretation is justified by the willingness of investigators to accept up to five false-positive conclusions in every 100 trials of regimens that, in truth, have equivalent efficacy. Unfortunately, even when a new treatment truly provides no advantage over a standard therapy, performing repeated analyses can greatly increase the chance of obtaining positive conclusions when this $p \leq 0.05$ guideline is repeatedly used.

As such, interim data safety reports pose well-recognized statistical problems related to the multiplicity of statistical tests to be conducted on the accumulating set of data. The basic problem is well known and is referred to as "sampling to a foregone conclusion" (Cornfield, 1966) and has been illustrated mathematically, pictorially or through simulations by many researchers (Fleming and Green, 1984). Specifically, in a simulation of 100 typical clinical trials of two interventions with truly equivalent efficacy that called for up to four periodic evaluations, 17 (rather than five) trials yielded false-positive conclusions (i.e. $p \leq 0.05$) in at least one analysis. The rate of false-positives continues to rise as the frequency of interim analyses rises. This serious increase in the likelihood of reaching false-positive conclusions due to misinterpretation of the strength of evidence when repeated analyses are conducted over time partly explains why many published claims of therapeutic advances have been false leads and provides the motivation for development of GSD.

A GSD first provides a schedule that relates patient accrual to when the interim analyses will occur. This schedule is conveniently expressed in terms of the proportion of the maximal possible number of patients that the trial could accrue. Second, such designs give a sequence of statistics used to test the null hypothesis, and third, they give a stopping rule defined in terms of a monotone increasing sequence of nominal significance levels at which each test will be conducted. This sequence of significance levels is carefully chosen to maintain the overall type I error at some desired level (e.g., 0.05 or 0.10) using one- or two-sided hypothesis. Either the number or the time of analyses is prespecified or the rate at which the overall significance level is "used up" is fixed in advance. Thus, undertaking group sequential trials assumes that hypothesis testing at nominal significance levels less than a prestated overall significance level will be performed, and that if results are ever extreme enough to exceed prespecified thresholds, the trial should be stopped. While such group sequential procedures differ in detail, they have certain common features.

The two commonly discussed pioneering mechanisms in GSD are given by Pocock (Pocock, 1977) and O'Brien and Fleming (OBF) (O'Brien and Fleming, 1979). Pocock adapted the idea of a repeated significance test at a constant nominal significance level to analyze accumulating data at a relatively small number of times over the course of the study. Patient entry was divided into equally sized groups and the data are analyzed after each group of observations has been collected. As an alternative, OBF proposed a test in which the nominal significance levels needed to reject the null hypothesis at sequential analyses increase as the study progresses, thus, making it more difficult to reject the null hypothesis at the earliest analysis but easier later on. Other variations to these schemes have also been developed but OBF is the most commonly utilized GSD as it fits well with the wishes of clinical trialists who do not want to stop a trial prematurely with insufficient evidence based on less reliable or unrepresentative data. There are other reasons for this preference. Historically, most clinical trials fail to show a significant treatment difference, hence from a global perspective, it is more cost-effective to use conservative designs. Indeed, even a conservative design such as OBF often shows a dramatic reduction in the average sample number (ASN or expected sample size) under the alternative hypothesis, H_A, compared to a FSD (see Table 1 for brief overview). Moreover, psychologically, it is preferable to have a nominal p-value at the end of the study for rejecting the null hypothesis, H_0, which is close to 0.05 in order to avoid the embarrassing situation where, say, a p-value of 0.03 at the final analysis would be declared non-significant.

Table 1
General properties of monitoring designs

Design	General	ASN (under H_0)	ASN (under H_A)
Fixed	Most conservative	Low	Large
OBF	Conservative, hard to stop early	Mid	Mid
Pocock	Most liberal, early stopping properties	Large	Low

Later, Wang et al. (1987) proposed a class of generalized formulation that encompasses Pocock and OBF methods as two extreme members.

Although the formulation of GSD started with binary outcomes, a generalized formulation has helped establish the wide applicability of the large sample theory for multivariate normal random variables with independent increments (i.e., standardized partial sums) to group sequential testing (Jennison and Turnbull, 1997; Scharfstein et al., 1997). This structure applies to the limiting distribution of test statistics which are fully efficient in parametric and semiparametric models, including generalized linear models and proportional hazards models (Tsiatis et al., 1995). It applies to all normal linear models, including mixed-effects models (Lee and Demets, 1991; Reboussin et al., 1992). Gange and Demets showed its applicability to the generalized estimating equation setting and Mazumdar and Liu showed the derivation for the comparative diagnostic test setting where area under the receiver operating characteristic curve is the endpoint (Mazumdar and Liu, 2003; Mazumdar, 2004). In short, almost any statistic likely to be used to summarize treatment differences in a clinical trial will justify group sequential testing with this basic structure and common mathematical formulation (Jennison and Turnbull, 2000).

3.1. Power and sample size calculation using inflation factor

Sample size computation in GSD setting involves the size of the treatment effect under some non-null hypothesis, the standard error of the estimated treatment effect at the end of the trial, and the drift of the underlying Brownian motion used to model the sequentially computed test statistics. The appropriate drift is determined by multiple factors such as the group sequential boundaries, type I error, and desired power. The theoretical background for design of group sequential trials has been discussed elsewhere (Kim and DeMets, 1992; Lan and Zucker, 1993) but the drift of commonly used GSDs can be easily translated into the corresponding IFs, provided in Table 2. The sample size approximation for a GSD in any setting is simply obtained by multiplying the sample size under the corresponding FSD by the IF provided in this table for the features of the specific GSD chosen. It is easy to note that the sample size inflation under OBF is minimal.

3.2. Monitoring boundaries using alpha spending functions

The earlier publications for group sequential boundaries required that the number and timing of interim analyses be fixed in advance. However, while monitoring data for real clinical trials, it was felt that more flexibility in being able to look at the data at time points dictated by the emerging beneficial or harmful trend is desired. To accommodate this capability, Lan and Demets proposed a more flexible implementation of the group sequential boundaries through an innovative 'alpha spending function' (Lan and Demets, 1983; Lan and DeMets, 1989). The spending function controls how much of the false-positive error (or false-negative error when testing to rule out benefit) can be used at each interim analysis as a function of the proportion (t^*, range 0 (study start)–1 (study

Table 2
Inflation Factors for Pocock and O'Brien–Fleming alpha spending functions for different total numbers of looks (K) under equal-sized increments

$\alpha = 0.05$ (Two-sided)					$\alpha = 0.01$ (Two-sided)				
K	Spending function	Power ($1-\beta$)			K	Spending function	Power ($1-\beta$)		
		0.80	0.90	0.95			0.80	0.90	0.95
2	Pocock	1.11	1.10	1.09	2	Pocock	1.09	1.08	1.08
2	OBF	1.01	1.01	1.01	2	OBF	1.00	1.00	1.00
3	Pocock	1.17	1.15	1.14	3	Pocock	1.14	1.12	1.12
3	OBF	1.02	1.02	1.02	3	OBF	1.01	1.01	1.01
4	Pocock	1.20	1.18	1.17	4	Pocock	1.17	1.15	1.14
4	OBF	1.02	1.02	1.02	4	OBF	1.01	1.01	1.01
5	Pocock	1.23	1.21	1.19	5	Pocock	1.19	1.17	1.16
5	OBF	1.03	1.03	1.02	5	OBF	1.02	1.01	1.01

end)) of total information observed. In many applications, t^* may be estimated as the fraction of patients recruited (for dichotomous outcomes) or the fraction of events observed (for time to event outcomes) out of the respective total expected. The alpha spending functions underlying OBF GSD correspond to

$$\alpha_1(t^*) = 2 - 2\Phi\left[\frac{Z_{1-(\alpha/2)}}{(t^*)^{1/2}}\right],$$

whereas the one for Pocock is described by

$$\alpha_2(t^*) = \alpha \ln[1 + (e-1)t^*].$$

The advantage of the alpha spending function is that neither the number nor the exact timing of the interim analyses needs to be specified in advance. Only the particular spending function needs to be specified. It is useful to note that the nominal significance levels utilized in any GSD will always add up to more than the overall significance level, because with multiple significance testing the probability of rejecting the null hypothesis does not accumulate additively due to positive correlations among test statistics.

Following is a sample 'Biostatistical Consideration' write-up for a clinical trial in Germ Cell Tumor (GCT) utilizing GSD with OBF boundaries. IF approach with three total looks ($K = 3$) was chosen at design stage and a series of boundaries and sequence of significance level were computed accordingly. The option of utilizing spending function approach was also kept open, which is often the case in practice.

3.3. Design of a phase 3 study with OBF GSD: A sample template

3.3.1. Biostatistical considerations

1. *Objective and background*: The objective of this study is to compare in a prospective randomized manner the efficacy of an experimental combination

regimen versus the standard regimen in previously untreated 'poor' risk GCT patients. The poor risk criterion helps identify patients who are expected to have high probability of worse outcome. It is described in the protocol and roughly depends on the primary site, histology, and specific blood markers being high. For this kind of cancer, a patient's prognosis is considered to be favorable if their tumor completely disappears and does not come back at least for a year. The response of these patients is called durable complete responder (DCR) at one year. In the institutional database at Memorial Sloan–Kettering Cancer Center (MSKCC) of size 796 patients treated by standard therapy, the proportion of patients remaining DCR at one year for the poor risk group ($n = 141$) is 30% with a 95% confidence interval (Cl) of 22.2–37.3%.

2. *Primary endpoint, power and significance level*: The major endpoint for this trial is DCR at one year where the time is computed from the day a patient is defined responder. This study is planned to detect a 20% absolute difference from the currently observed rate of 30% (30% versus 50%). We are expecting an accrual of 50 patients per year. The sample size calculation based on log-rank test for an FSD with 80% power and 5% level of significance, 195 patients will be needed. To incorporate two interim looks and a final look (so total $K = 3$) at the end of full accrual, an IF of 1.02 was multiplied to 195 requiring 199 patients ($= 1.02 \times 195$) using OBF method (O'Brien and Fleming 1979). Rounding it off to 200 patients (100 per arm), we decide to place the two interim looks at the end of second and third year and the final look at the end of fourth year as the accrual rate of 50 patients makes the length of study to be four years.
3. *Randomization*: After eligibility is established, patients will be randomized via a telephone call to the coordinating center at MSKCC clinical trial office (Phone number: XXX-XX-XXXX; 9:00 am to 5:00pm Monday through Friday). Randomization will be accomplished by the method of stratified random permuted block, where patient institution (MSKCC versus ECOG versus SWOG versus remaining participating institutions) was adopted for stratification, where ECOG denotes Eastern Cooperative Oncology Group and SWOG denotes Southwest Oncology Group.
4. *Data safety monitoring committee and interim analyses*: The data will be reviewed at designated intervals by an independent DSMC. This committee was formed with two independent oncologists and one independent biostatistician. The committee will be presented with the data summary on accrual rates, demographics and bio-chemical markers etc. and comparative analysis (using Fisher's exact test) on toxicity and DCR proportion by the principal investigator (PI) and the biostatistician on study. Survival and progression-free survival curves will be estimated only if there is an enough number of events that governs statistical power. Semi-annual reports on toxicity will be disseminated to all the participating groups.

Normalized z-statistics according to the OBF boundary to be used for stopping early if the experimental regimen looks promising are ± 3.471, ± 2.454, ± 2.004, where the corresponding sequence of nominal significance levels are 0.001, 0.014, and 0.036, respectively (East, Cytel Statistical Software). If situation emerges where these time points are not the most convenient or desirable, Lan–Demets spending function utilizing OBF boundaries will be used to compute the corresponding z-statistics and significance level. The committee is expected to use the statistical stopping rules as a guideline in addition to both medical judgment and the relevant emerging data in the literature, especially ones obtained from similar trials.

5. *Final analysis*: All toxicities will be evaluated based on the NCI common toxicity criteria and tabulated by their frequencies and proportions. Fisher's exact test will be used to compare the toxicities and adverse events by the two arms. The primary analysis, DCR-free survival curves will be estimated using Kaplan–Meier method and with appropriate follow-up, comparisons will be made using log-rank test (Kaplan and Meier, 1958; Mantel, 1966). Once the trial stops (either at interim look or at final look), standard statistical estimation and inference will be undertaken for the observed treatment difference.

3.4. *Analyses following group sequential test*

Analysis following a group sequential test consists of two scenarios: The first is upon conclusion of the trial after the test statistic has crossed a stopping boundary and the second is when an interval estimate of the treatment difference is desired whether the design calls for a termination or not. Tsiatis et al. (1984) have shown that in both situation, it is inappropriate to compute a 'naïve' CI, treating the data as if they had been obtained in a fixed sample size experiment. They estimated naïve CI following a five-stage Pocock's test with 5% level of significance and found their coverage to vary between 84.6% and 92.9%, depending on the true parameter value.

For the first scenario, Tsiatis et al. suggested a numerical method for calculating an exact CIs following group sequential tests with Pocock (1977) or O'Brien and Fleming (1979) boundaries based on ordering the sample space in a specific manner. They derived the CIs based on normal distribution theory, which pull the naive CIs toward zero and are no longer symmetric about the sample mean. They also commented that their method is applicable to any (asymptotically) normal test statistic which has uncorrelated increments and for which the variance can be estimated consistently. Whitehead (1986) suggested an approach for adjusting the maximum likelihood estimate as the point estimate by subtracting an estimate of the bias. Wang and Leung (1997) proposed a parametric bootstrap method for finding a bias-adjusted estimate, whereas Emerson and Fleming (1990) provide a formulation of uniformly minimum variance unbiased estimator calculated by Rao–Blackwell technique.

For the second scenario, the multiple-looks problem affects the construction of CIs just as it affects significance levels of hypothesis tests. Repeated CIs for a parameter θ are defined as a sequence of intervals I_k, $k = 1, \ldots, K$, for which a simultaneous coverage probability is maintained at some level, say, $1-\alpha$. The defining property of a $(1-\alpha)$-level sequence of repeated CIs for θ is $P[\theta \in I_k \text{ for all } k = 1, \ldots, K] = 1-\alpha$ for all θ (Jennison and Turnbull, 1983, 1984, 1985). The interval I_k, $k = 1, \ldots, K$, provides a statistical summary of the information about the parameter θ at the kth analysis, automatically adjusted to compensate for repeated looks at the accumulating data. As a result, repeated CIs instead of group sequential testing can be used for monitoring clinical trials (Jennison and Turnbull, 1989).

Most conventional trials are designed to have a high probability of detecting a predefined treatment effect if such an effect truly exists. That probability is called the power of the trial. Most trials use power in the range of 0.8–0.95 for a plausible range of alternatives of interest and the sample size of the study is calculated to achieve that power. The concept of 'conditional power' comes into play when supporting evidence is sought to decide the power midstream.

3.5. Stochastic curtailment

Once the trial starts and data become available, the probability that a treatment effect will ultimately be detected can be recalculated (Halperin et al., 1982; Lan et al., 1982; Lan and Wittes, 1988). An emerging trend in favor of the treatment increases the probability that the trial will detect a beneficial effect, while an unfavorable trend decreases the probability of establishing benefit. The term 'conditional power' is often used to describe this evolving probability. The term 'power' is used because it is the probability of claiming a treatment difference at the end of the trial, but it is 'conditional' because it takes into consideration the data already observed that will be part of the final analysis. Conditional power can be calculated for a variety of scenarios including a positive beneficial trend, a negative harmful trend, or no trend at all. However, these calculations are frequently made when interim data are viewed to be unfavorable. For this scenario, it represents the probability that the current unfavorable trend would improve sufficiently to yield statistically significant evidence of benefit by the scheduled end of the trial. This probability is usually computed under the assumption that the remainder of the data will be generated from a setting in which the true treatment effect was as large as the originally hypothesized in the study protocol.

When an unfavorable trend is observed at the interim analysis, the conditional probability of achieving a statistically significant beneficial effect is much less than the initial power of the trial. If the conditional power is low for a wide range of reasonable assumed treatment effect, including those originally assumed in the protocol, this might suggest to the DSMC that there is little reason to continue the trial since the treatment is highly unlikely to show benefit. Of course, this conditional power calculation does increase the chance of missing a real benefit (false-negative or type II error) since termination eliminates any chance of recovery by the intervention. However, if the conditional power under these scenarios is less than 0.2 compared to the hypothesis for which the trial originally provided power of 0.85–0.9, the increase in the rate of false-negative error is negligible. There is no concern with false-positive error in this situation since there is no consideration of claiming a positive result. An example of its use will follow in the Beta-Blocker Heart Attack Trial (BHAT) trial description later in this chapter.

3.6. Bayesian monitoring

The Bayesian approach for monitoring accumulating data considers unknown parameters to be random and to follow probability distributions (Spiegelhalter et al., 1986; Freedman et al., 1994; Parmar et al., 1994; Fayers et al., 1997). The investigators specify a prior distribution(s) describing the uncertainty in the treatment effect and other relevant parameters. These prior distributions are developed based on previous data and beliefs. It is quantified through a distribution of possible values and is referred to as the prior distribution. The observed accumulating data are used to modify the prior distribution and produce a posterior distribution, a distribution that reflects the most current information on the treatment effect, taking into account the specified prior as well as the accumulated data. This posterior distribution can then be used to compute a variety of summaries including the predictive probability that the treatment is effective. In 1966, Cornfield introduced the idea of Bayesian approach to monitoring clinical trial (Cornfield, 1966). Although, interest has recently increased in its use (Kpozehouen et al., 2005) and availability of computational tools have made it more feasible to use, these methods are still not widely utilized.

3.7. Available softwares

Softwares for implementing GSDs have been developed and commercialized since the early 1990s. Extended descriptions of these softwares are available through their user's guide and some review papers (Emerson, 1996; Wassmer and Vandemeulebroecke, 2006). Most of the computational tools employ the recursive numerical integration technique that takes advantage of a quadrature rule of replacing integral by a weighted sum for probabilistic computations (Armitage et al., 1969; Jennison and Turnbull, 2000).

Here, we provide a comprehensive listing of appropriate links for free self-executable softwares as well as codes written in FORTRAN, SAS, Splus, and R languages. FORTRAN source code used in the textbook by Jennison and Turnbull (2000) can be downloaded from Dr. Jennison's homepage on http://people.bath.ac.uk/mascj/book/programs/general. The code provides continuation regions and exit probabilities for classical GSDs including those proposed by Pocock (1977), O'Brien and Fleming (1979), Wang and Tsiatis (1987) and Pampallona and Tsiatis (1994). In addition, the spending function approach according to Lan and Demets (1983) is implemented. Another implementation in FORTRAN of the spending function approach is available for use under UNIX and MS-DOS. It can be downloaded from http://www.biostat.wisc.edu/landemets/ as a stand-alone program with a graphical user interface, while details of methodologies and algorithms are found in Reboussin et al. (2000). These codes provide computation of boundaries and exit probabilities for any trial based on normally or asymptotic normally distributed test statistics with independent increments, including those in which patients give a single continuous or binary response, survival studies, and certain longitudinal designs. Interim analyses need not be equally spaced, and their number need not be specified in advance via flexible alpha spending mechanism. In addition to boundaries, power computations, probabilities associated with a given set of boundaries, and CIs can also be computed.

The IML (Interactive Matrix Language) module of SAS® features the calls SEQ, SEQSCALE, and SEQSHIFT that perform computations for group sequential tests. SEQ calculates the exit probabilities for a set of successive continuation intervals. SEQSCALE scales these continuation regions to achieve a specified overall significance level and also returns the corresponding exit probabilities. SEQSHIFT computes the non-centrality parameter for a given power.

S-PLUS that is commercially available provides a package for designing, monitoring, and analyzing group sequential trials through its S + SeqTrial™ module. It makes use of the unifying formulation by Kittelson et al. (Kittelson and Emerson, 1999), including all classical GSDs, triangular tests (Whitehead, 1997), and the spending function approach. It offers the calculation of continuation regions, exit probabilities, power, sample size distributions, overall *p*-values and adjusted point estimates and CIs, for a variety of distributional assumptions. It comes with a graphical user interface and very good documentation, which can be downloaded from http://www.insightful.com/products/seqtrial/default.asp.

In R (http://www.r-project.org/), cumulative exit probabilities of GSDs can be computed by the function seqmon. It implements an algorithm proposed by Schoenfeld (2001) and the documentation and packages are freely downloadable at http://www.maths.lth.se/help/R/.R/library/seqmon/html/seqmon.html.

PEST, version 4 offers a wide range of scenarios, including binary, normal, and survival endpoints, and different types of design. The main focus of PEST is the implementation of triangular designs. Sequential designs from outside PEST can also be entered and analyzed. Besides the planning tools, the software offers a number of analysis tools including interim monitoring and adjusted *p*-values, CIs, and point estimates for the final analysis. An important and unique feature of PEST is that interim and final data can be optionally read from SAS data sets. More information about the software can be found at http://www.rdg.ac.uk/mps/mps_home/software/software.htm#PEST%204.

East of Cytel Statistical Software and Services (http://www.cytel.com/Products/East/) is the most comprehensive package for planning and analyzing group sequential trials. The software provides a variety of capabilities of advanced clinical trial design, simulation and monitoring, and comes with extensive documentation including many real data examples. Tutorial sessions for East are frequently offered during various statistical meetings and conferences and educational settings.

"PASS 2005 Power Analysis and Sample Size" is distributed by NCSS Inc. This software supplies the critical regions and the necessary sample sizes but it is not yet possible to apply a sequential test to real data in the sense of performing an adjusted analysis (point estimates, CIs, and *p*-values). Documentation and a free download are available on http://www.ncss.com./passsequence.html.

"ADDPLAN Adaptive Designs-Plans and Analyses" (http://www.addplan.com/) is designed for the purpose of planning and conducting a clinical trial based on an adaptive group sequential test design. New adaptive (flexible) study designs allow for correct data-driven re-estimation of the sample size while controlling the type I error rate. Redesigning the sample size in an interim analysis based on the results observed so far considerably improves the power of the trial since the best available information at hand is used for the sample size adjustment. The simulation capabilities for specific adaptation rules are also provided.

The choice of software is based on the users' need and the complexity of design. The freely available softwares are often enough to implement basic functions to be used in standard or popular designs and to perform associated data analyses outlined in this chapter unless special features are required.

3.8. Data safety monitoring committee

Early in the development of modern clinical trial methodology, some investigators recognized that, despite the compelling ethical needs to monitor the accumulating results, repeated review of interim data raised some problems. It was recognized that knowledge of the pattern of the accumulating data on the part of investigators, sponsors, or trial participants, could affect the course of the trial and the validity of the results. For example, if investigators were aware that the interim trial results were favoring one of the treatment groups, they might be reluctant to continue to encourage adherence to all regimens in the trial, or to continue to enter patients in the trial, or they may alter the types of patients they would consider accrual. Furthermore, influenced by financial or scientific conflicts of interest, investigators, or the sponsor might take actions that could diminish the integrity and credibility of the trial. A natural and practical approach to dealing with this problem is to assign sole responsibility for interim monitoring of data on safety and efficacy to a committee whose members have no involvement in the trial, no vested interest in the trial results, and sufficient understanding of the trial design, conduct, and data-analytical issues to interpret interim analyses with appropriate caution. These DSMCs consisting of members from variety of background (clinical, statistical, ethical, etc.) have become critical components of virtually all clinical trials.

For the above example, an independent DSMC consisting of three members with background in oncology (one from community hospital and one from specialized center) and biostatistics met every year to discuss the progress of the trial. The outcome comparison was only presented when an interim analysis with OBF was allowed. Below we present a list of items that were included in the interim report for this trial. This is a typical template for a clinical trial and could be useful in other scenarios.

Items included in the interim report:

1. Brief outline of the study design
2. Major protocol amendments with dates (or summary) if applicable
3. Enrollment by arm and year and center (preferably, updated within a month of the DSMC meeting date)
4. Information on eligibility criterion violation or crossover patients
5. Summary statistics (e.g., mean/median) on follow-up times of patients
6. Frequency tables of baseline characteristics (demographics, toxicity, and

adverse event summary, laboratory test summary, precious treatment) of the full cohort

7. Comparative analysis of primary and secondary endpoints (when data mature)
8. Subgroup analyses and analyses adjusted for baseline characteristics (and some secondary outcomes data, if any)
9. Comparative analysis of adverse event and toxicity data
10. Comparative analysis of longitudinal lab values.

The GCT study referred above struggled with accrual of patients and remained open for 10 years instead of the four years planned initially. To improve accrual rate, new centers were added and the patient eligibility was expanded. DSMC met annually and approved these actions. The first DSMC meeting where outcome data were compared was at 6th year after study start instead of the 2nd year. Lan–Demets with OBF boundary was utilized to compute the appropriate boundary but the boundary was not crossed. DSMC deliberations continued with concern for the accrual rate but since the experimental regimen utilizing autologus bone marrow transplant was quite a novel and unique approach and it was added to the standard therapy, the DSMC did not feel any harm to patients and decided to keep the trial open. More assertive accrual plans were adopted but when many of these plans failed to improve accrual, the study was at last closed at 219 patients (in contrast, $N = 270$ in the original plan).

3.8.1. Details included in the final paper (on design and primary analysis)

The final write-up or summary report needs to include as much details as possible about the original design (including sample size/power calculation), modifications, rationale for modification, decisions by DSMC, and conclusions. Here's part of the 'Statistical Methods' section from the final paper related to the GCT study (Motzer et al., 2007):

> The trial was designed with the proportion of patients with durable complete response (DCR) at one year from entry onto the trial as the primary endpoint. The original study population to be enrolled on this study was poor-risk GCT patients only. We had planned to accrue 200 patients (100 per arm) to detect a 20% difference in DCR rate at one year (an improvement from 30% to 50%) with a 5% level of significance and 80% power. However, as the trial progressed, the accrual rate was far lower than our expectation of 50 poor-risk patients per year. Also during this time, an international effort brought along a newly developed but broadly accepted risk group classification and it was felt that the intermediate-risk group patients with poor markers (lactate dehydrogenase greater than 3 times upper limit of normal) would benefit from the treatment under investigation. Therefore it was decided to extend the study to this modified intermediate risk group from the poor risk classification utilized before. Based on a historical one-year DCR rate of 45% in the poor and intermediate risk groups combined, we then modified our target accrual to 218 patients to detect an improvement of 20% with the same level and power.

A final modification to the study was implemented in 2002 after a new center CALGB was added to the study and accrual at that center began. At that point, it was our hope to be able to address the original question of interest in the poor-risk group of patients. We planned to accrue 270 patients, consisting of 216 poor-risk patients (200 per original calculation +16 to account for withdrawals) and 54 intermediate-risk patients. However, as accrual did not meet our expectations even with the additional cooperative group participating, the study was closed in August of 2003. The data were reviewed annually by an independent DSMC. Initially, the design included an O'Brien and Fleming stopping rule with the sequence of nominal significance levels of 0.001, 0.014, and 0.036 for the two interim analyses and the final analysis, respectively. A formal comparative interim analysis on DCR proportion and overall survival was presented in May 2000 based on a recalculated boundary utilizing Lan–Demets spending function. The decision was to continue the trial as the boundary was not crossed and no ethical conflict was found since the experimental regimen was an autologus bone marrow transplant regimen on top of the standard therapy. The study was at last stopped in 2003 due to not being able to improve accrual rate.

3.9. Historical example of GSD use

It is always educational to look back on the trials that were planned with GSD and benefited from it. Two excellent books by DeMets et al., 2006 and Ellenberg et al., 2006 provide essential and in-depth reading materials for clinical trialists starting in this field. An example considered by these books and many other publications is described below to show the multifaceted decision process that goes into the deliberation of DSMB.

The BHAT compared the beta-blocker propranolol against placebo in patients who had a myocardial infarction recently. The statistical design called for enrollment of 4,020 patients, aged 30–69 years, who had a myocardial infarction 5–21 days prior to randomization. The primary objective of the study was to determine if long-term administration of propranolol would result in a difference in all-cause mortality. The design utilized O'Brien–Fleming boundary with alpha level set at two-tailed 0.05, 90% power, and three-year average follow-up. The attempt was to detect a 21.25% relative change in mortality, from a three-year rate of 17.46% in the control (placebo) group to 13.75% in the intervention group, which were obtained from earlier studies (Furberg and Friedwald, 1978; Anderson et al., 1979) after taking non-adherence into account (Byington, 1984).

Enrollment began in 1978 and a total of 3,837 participants were accrued instead of the planned 4,020. This reduced the power slightly from the planned 90% to 89%. The PDMB first reviewed the data in May 1979. Subsequent data reviews were to occur approximately every six months, until the scheduled end of the trial in June 1982. At the *October, 1979* meeting of the PDMB, the log-rank z-value exceeded the conventional 1.96 critical value for a nominal p of 0.05 but was far from significance due to the conservative nature of the O'Brien–Fleming boundaries early in the study. PDMB recommended continuation of the trial.

At the meeting in *April 1981*, the PDMB reviewed not only the accumulating BHAT data but the results of the timolol trial that had just been published. This trial of 1,884 survivors of an acute myocardial infarction showed a statistically significant reduction in all-cause mortality, from 16.2% to 10.4%, during a mean follow-up of 17 months. At this point, BHAT was no longer enrolling patients, but follow-up was continuing. The PDMB recommended that BHAT continues, primarily because, despite the timolol findings, the BHAT data did not show convincing evidence of benefit. Not only had the monitoring boundary not been crossed, but the long-term effect on mortality and possible adverse events was unknown. Importantly, all patients in BHAT had been in the trial for at least six months post-infarction, and there was no evidence that beta-blockers started after that time produced benefit. Thus, there was not an ethical concern about leaving the participants on placebo. The PDMB advised that the study investigators be informed of the timolol results. However, it also advised that because there had been conflicting results from other beta-blocker trials, the positive results of the timolol trial should not preclude the continuation of BHAT. Furthermore, timolol was not available for sale in the United States then. At its *October 1981* data review, the PDMB noted that the upper OBF boundary had been crossed. The normalized log-rank statistic was then 2.82, which exceeded the boundary value of 2.23. In addition to the monitoring boundaries, the PDMB considered a number of factors in its recommendation to stop early:

> 1) Conditional power calculations indicated that there was little likelihood that the conclusions of the study would be changed if follow-up were to continue; 2) The gain in precision of the estimated results for the first two years would be tiny, and only modest for the third year; 3) The results were consistent with those of another beta-blocker trial; 4) There would be potential medical benefits to both study participants on placebo and to heart attack patients outside the study; 5) Other characteristics, such as subgroup examinations and baseline comparability, confirmed the validity of the findings; 6) The consent form clearly called for the study to end when benefit was known. Following points in favor of continuing until the scheduled end were considered but were not found to weigh enough in favor of not stopping: 1) Even though slight, there remained a chance that the conclusions could change; 2) Because therapy would be continued indefinitely, it would be important to obtain more long-term (4 year) data; 3) It would be important to obtain more data on subgroups and secondary outcomes; 4) The results of a study that stopped early would not be as persuasive to the medical community as would results from a fully powered study that went to completion, particularly given the mixed results from previous trials.

Lessons learnt from these experiences are that 1) O'Brien-Fleming approach to sequential boundaries could prove very helpful in fostering a cautious attitude with regard to claiming significance prematurely. Even though conventional significance was seen early in the study, the use of sequential boundaries gave the study added credibility and probably helped make it persuasive to the practicing medical community; 2) The use of conditional power added to the persuasiveness of the results, by showing the extremely low likelihood that the conclusions would change if the trial were to continue to its scheduled end; 3) The decision-making process involves many factors, only some of which are statistical (Friedman et al., 2003).

4. Steps for GSD design and analysis

4.1. Classical design

Step 1: Decide the number of maximum looks (or groups) K and the choice of boundary (that can be indexed by shape parameter, Δ (Wang and Tsiatis, 1987).

Remark:

a) The gain in ASN is most dramatic when going from $K = 1$ (i.e., the fixed sample size design) to $K = 2$. Beyond $K = 5$, there is relatively little change in ASN.

b) The choice of K may be dictated by some practicality such as the frequency of the DSMC meetings that is feasible.

c) $\Delta = 0$ for OBF and $\Delta = 0.5$ for Pocock.

Step 2: Compute the sample size for fixed design as you would ordinarily do (using significance level, power, and effect size). Multiply by the appropriate IF.

Step 3: After computing the maximum sample size, divide it into K equal group sizes and conduct interim analyses after each group. Reject H_0 at the first interim analysis where the test statistic using all the accumulated data exceeds the boundary values computed. Alternatively, we can translate the boundaries to the corresponding nominal p-values at each look and conduct the test using p-values.

4.2. Information-based design

Step 1: Specify level of significance, power, K and alternative of interest (γ).

Remark:

You specify K at the design stage but you may deviate from this at the time of analysis.

Step 2: Choose a spending function and stopping boundary (Lan and DeMets spending function with OBF or Pocock or other boundaries).

Step 3: Compute maximum information (MI) required to have a specific power as $\mathrm{MI} = (z_{1-\alpha/2} + z_{1-\beta}/\gamma)^2$ X IF.

Step 4: The first time the data are monitored, say, at time t_1, compute the proportion of information compared to MI. Then find the first boundary value. If the test statistic exceeds the boundary computed, stop and reject H_0. If not, continue to next monitoring time.

Step 5: At time t_2, compute the ratio of observed information and MI. Then perform the testing.

Step 6: Continue in this fashion, if necessary, until the final analysis, at which point you use up the remaining significance level.

Remark:

With this strategy, you are guaranteed a level alpha test regardless of how often or when you look at the data prior to obtaining MI.

5. Discussion

In RCTs designed to assess the efficacy and safety of medical interventions, evolving data are typically reviewed on a periodic basis during the conduct of the study. These interim reviews are especially important in trials conducted in the setting of diseases that are life-threatening or result in irreversible major morbidity. Such reviews have many purposes. They may identify unacceptably slow rates of accrual or high rates of ineligibility determined after randomization, protocol violations that suggest that clarification of or changes to the study protocol are needed or unexpectedly high dropout rates that threaten the trial's ability to produce unbiased results. The most important purpose, however, is to ensure that the trial remains appropriate and safe for the individuals who have been or are still to be enrolled. Efficacy results must also be monitored to enable benefit-to-risk assessments to be made. Repeated statistical testing of the primary efficacy endpoint was seen to increase the chance of false-positive rate. The methods of adjusting the significance levels at each interim analysis so that the overall false-positive rate stays at an acceptable level gave rise to GSDs. The field has been developing for past 30 years and is now quite mature with various methods with well-studied operating characteristics and availability of an array of user-friendly software.

One new field of applications has been cluster-randomized trials (CRTs). CRTs have been used increasingly over the past two decades to measure the effects of health interventions applied at the community level. Excellent reviews and books are written by Donner et al. and Murray (Donner and Brown, 1990; Murray, 1998; Donner and Klar, 2000). Recently, Zou et al. (2005) developed group sequential methods that can be applied to CRT. Although the design aspect is well characterized and related computer program is available upon request, effect estimation following this group sequential test remains a topic of future research. This method is not yet used prospectively on a clinical trial. Development of methodology for novel design such as the split-cluster design could also be a useful addition to this field (Donner and Klar, 2004).

Adaptive designs in the context of group sequential testing allow modifications of particular aspects of the trials (such as inappropriate assumptions, excessive cost, or saving in time) after its initiation without undermining the validity and integrity of the trial. Some developments have been made to combine the advantages of adaptive and of classical group sequential approaches. Although research has been ongoing in this field, it still remains a field of research priority (Tsiatis and Mehta, 2003; Jennison and Turnbull, 2005; Kuehn, 2006; Wassmer, 2006).

There are some settings where GSDs may not be appropriate. For example, when the endpoint assessment time is lengthy relative to the recruitment period, there might be enough interim results to perform an analysis only after all or most subjects have been recruited and treated, thereby potentially rendering the GSD irrelevant. Most other large studies will benefit from having planned look at the data as trial progresses. Quite surprisingly, we found that many large trials follow FSD (Cooper et al., 2006; Cotton et al., 2006; Nicholls et al., 2006). A systematic literature search to assess the percentage of studies that would benefit from GSD but is not currently planning to use it would be interesting. This effort could also identify additional areas for further research or need for expanded exposure of these designs among practitioners.

Acknowledgement

We thank Ms. Anita Mesi for her excellent help with the published literature management using endnote software and past collaborators, Dr. Robert Motzer, and Ms. Jennifer Bacik, for many discussions on this topic. Partial support for this work came from the following grants: CERTs (AHRQ RFA-HS-05-14), CIPRA (NIAID U01 AI058257), R25 CA105012, and Cornell Institute of Clinical Research (supported by Tolly Vinik Trust).

References

Anderson, M., Bechgaard, P., Frederiksen, J. (1979). Effect of Alprenolol on mortality among patients with definite or suspected acute myocardial infarction: Preliminary results. *Lancet* **2**, 865–868.

Anderson, T. (1960). A modification of the sequential probability ratio test to reduce the sample size. *Ann Math Stat* **31**, 165–197.

Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *Quarterly Journal of Medicine* **23**, 255–274.

Armitage, P., McPherson, C.K., Rowe, B.C. (1969). Repeated significance tests on accumulating data. *Journal of Royal Statistical Society. Series A* **132**, 235–244.

Betensky, R.A. (1997). Early stopping to accept H(o) based on conditional power: Approximations and comparisons. *Biometrics* **53**(3), 794–806.

Bross, I. (1952). Sequential medical plans. *Biometrics* **8**, 188–205.

Byington, R. (1984). Beta-Blocker Heart Attack Trial: Design, methods, and baseline results. *Controlled Clinical Trials* **5**, 382–437.

Canner, P.L. (1977). Monitoring treatment differences in long-term clinical trials. *Biometrics* **33**(4), 603–615.

Cooper, C.J., Murphy, T.P. et al. (2006). Stent revascularization for the prevention of cardiovascular and renal events among patients with renal artery stenosis and systolic hypertension: Rationale and design of the CORAL trial. *American Heart Journal* **152**(1), 59–66.

Cornfield, J. (1966). A Bayesian test of some classical hypotheses – with application to sequential clinical trials. *Journal of the American Statistical Association* **61**, 577–594.

Cotton, S.C., Sharp, L. et al. (2006). Trial of management of borderline and other low-grade abnormal smears (TOMBOLA): Trial design. *Contemporary Clinical Trials* **27**(5), 449–471.

DeMets, D., Furberg, C. et al. (2006). *Data Monitoring in Clinical Trials: A Case Studies Approach.* Springer, New York.

Donner, A., Brown, K. (1990). A methodological review of non-therapeutic inter vention trials employing cluster randomization. *International Journal of Epidemiology* **19**(4), 795–800.

Donner, A., Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research.* Arnold, London.

Donner, A., Klar, N. (2004). Methods for statistical analysis of binary data in split-cluster designs. *Biometrics* **60**(4), 919–925.

Elfring, G.L., Schultz, J.R. et al. (1973). Group sequential designs for clinical trials. *Biometrics* **29**(3), 471–477.

Ellenberg, S.S. (2001). Independent monitoring committees: Rationale, operations, and controversies. *Statistics in Medicine* **20**, 2573–2583.

Ellenberg, S.S., Fleming, T. et al. (2006). *Data Monitoring Committees in Clinical Trials: A Practical Perspective*. Wiley, London.

Emerson, S. (1996). Statistical packages for group sequential methods. *The American Statistician* **50**, 183–192.

Emerson, S., Fleming, T. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**, 875–892.

Fayers, P.M., Ashby, D. et al. (1997). Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Statistics in Medicine* **16**, 1413–1430.

Fleming, T., DeMets, D. (1993). Monitoring of clinical trials: Issues and recommendations. *Controlled Clinical Trials* **14**(3), 183–197.

Fleming, T.R., Green, S. (1984). Considerations for monitoring and evaluating treatment effect in clinical trials. *Controlled Clinical Trials* **5**, 55–66.

Fleming, T.R., Watelet, L.F. (1989). Approaches to monitoring clinical trials. *Journal of the National Cancer Institute* **81**, 188–193.

Freedman, L., Spiegelhalter, D. et al. (1994). The what, why, and how of Bayesian clinical trials monitoring. *Statistics in Medicine* **13**, 1371–1383.

Friedman, L., Demets, D., et al. (2003). Data and safety monitoring in the Beta-Blocker Heart Attach Trial: Early experience in formal monitoring methods.

Frustaci, S., Gherlinzoni, F. et al. (2001). Adjuvant chemotherapy for adult soft tissue sarcomas of the extremities and girdles: Results of the Italian randomized cooperative trial. *Journal of Clinical Oncology* **19**, 1238–1247.

Furberg, C., Friedwald, W. (Eds.) (1978). Effects of chronic administration of beta-blockade on long-term survival following myocardial infarction. *Beta-Adrenergic Blockade: A New Era in Cardiovascular Medicine.* Excerpta Medica, Amsterdam.

Gausche, M., Lewis, R.J. et al. (2000). Effect of out-of-hospital pediatric endotracheal intubation on survival and neurological outcome. *The Journal of the American Medical Association* **283**(6), 783–790.

Geller, N.L., Pocock, S.J. et al. (1987). Interim analyses in randomized clinical trials: Ramifications and guidelines for practitioners. *Biometrics* **43**(1), 213–223.

Halperin, M., Lan, K. et al. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials* **3**, 311–323.

Jennison, D., Turnbull, B. (1983). Confidence interval for a bionomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* **25**, 49–63.

Jennison, C., Turnbull, B. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* **5**, 33–45.

Jennison, C., Turnbull, B. (1985). Repeated confidence intervals for the median survival time. *Biometrika* **72**, 619–625.

Jennison, C., Turnbull, B. (1989). Interim Analyses: The repeated confidence interval approach (with discussion). *Journal of Royal Statistical Society. Series B* **51**, 305–361.

Jennison, C., Turnbull, B.W. (1997). Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92**, 1330–1341.

Jennison, C., Turnbull, B.W. (2000). *Group Sequential Methods with Application to Clinical Trials.* Chapman & Hall.

Jennison, C., Turnbull, B.W. et al. (2005). Meta-analyses and adaptive group sequential designs in the clinical development process. *Journal of Biopharmaceutical Statistics* **15**(4), 537–558.

Jones, D., Newman, C. et al. (1982). The design of a sequential clinical trial for the comparison of two lung cancer treatments. *Statistics in Medicine* **1**(1), 73–82.

Kaplan, E., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association* **53**, 457–481.

Kelly, K., Crowley, J. et al. (2001). Randomized phase III trial of paclitaxel plus carboplatin versus vinorelbine plus cisplatin in the treatment of patients with advanced non–small-cell lung cancer: A southwest oncology group trial. *Journal of Clinical Oncology* **19**, 3210–3218.

Kim, K., DeMets, D. (1992). Sample size determination for group sequential clinical trials with immediate response. *Statistics in Medicine* **11**(10), 1391–1399.

Kittelson, J., Emerson, S. (1999). A unifying family of group sequential test designs. *Biometrics* **55**, 874–882.

Kpozehouen, A., Alioum, A. et al. (2005). Use of a Bayesian approach to decide when to stop a therapeutic trial: The case of a chemoprophylaxis trial in human immunodeficiency virus infection. *American Journal of Epidemiology,* **161**(6), 595–603, (see comment).

Kuehn, B. (2006). Industry, FDA warm to "Adaptive" trials. *The Journal of the American Medical Association* **296**(16), 1955–1971.

Lan, K., DeMets, D.L. (1989). Group sequential procedures: Calendar versus information time. *Statistics in Medicine* **8**, 1191–1198.

Lan, K., Demets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

Lan, K., Simon, R. et al. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics C* **1**, 207–219.

Lan, K., Wittes, J. (1988). The B-value: A tool for monitoring data. *Biometrics* **44**, 579–585.

Lan, K., Zucker, D. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion. *Statistics in Medicine* **12**, 753–765.

Lee, J., Demets, D. (1991). Sequential comparison of changes with repeated measurement data. *Journal of American Statistical Association* **86**, 757–762.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163–170.

Mazumdar, M. (2004). Group sequential design for comparative diagnostic accuracy studies: Implications and guidelines for practitioners. *Medical Decision Making: An International Journal of the Society for Medical Decision Making* **24**(5), 525–533.

Mazumdar, M., Liu, A. (2003). Group sequential design for comparative diagnostic accuracy studies. *Statistics in Medicine* **22**(5), 727–739.

McPherson, K. (1974). Statistics: The problem of examining accumulating data more than once. *New England Journal of Medicine* **290**, 501–502.

Motzer, R., Nichols, C. et al. (2007). Phase III randomized trial of conventional-dose chemotherapy with or without high-dose chemotherapy and autologous hematopoietic stem-cell rescue as first-line treatment for patients with poor-prognosis metastatic germ cell tumors. *Journal of Clinical Oncology* **25**(3), 247–256.

Murray, D.M. (1998). *Design and Analysis of Group-randomized Trials.* Oxford University Press, New York.

Nicholls, S.J., Sipahi, I. et al. (2006). Intravascular ultrasound assessment of novel antiatherosclerotic therapies: Rationale and design of the Acyl-CoA:Cholesterol Acyltransferase Intravascular Atherosclerosis Treatment Evaluation (ACTIVATE) Study. *American Heart Journal* **152**(1), 67–74.

O'Brien, P., Fleming, T. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Pampallona, S., Tsiatis, A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of null hypothesis. *Journal of Statistical Planning and Inference* **42**, 19–35.

Parmar, M., Spiegelhalter, D. et al. (1994). The CHART trials: Bayesian design and monitoring in practice. *Statistics in Medicine* **13**, 1297–1312.

Pepe, M., Anderson, G. (1992). Two-stage experimental designs: Early stopping with a negative result. *Applied Statistics* **41**(1), 181–190.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

Proschan, M., Lan, K. et al. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach.* Springer.

Reboussin, D., Lan, K. et al. (1992). Group Sequential Testing of Longitudinal Data. Tech Report No. 72, Department of Biostatistics, University of Wisconsin.

Reboussin, D.M., DeMets, D.L. et al. (2000). Computations for group sequential boundaries using the Lan–DeMets spending function method. *Controlled Clinical Trials* **21**(3), 190–207.

Sacco, R.L., DeRosa, J.T. et al. (2001). Glycine antagonist in neuroprotection for patients with acute stroke: GAIN Americas – a randomized controlled trial. *The Journal of the American Medical Association* **285**(13), 1719–1728.

Scharfstein, D., Tsiatis, A. et al. (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* **92**, 1342–1350.

Schoenfeld, D.A. (2001). A simple algorithm for designing group sequential clinical trials. *Biometrics* **57**(3), 972–974.

Sebille, V., Bellissant, E. (2003). Sequential methods and group sequential designs for comparative clinical trials. *Fundamental and Clinical Pharmacology* **17**(5), 505–516.

Spiegelhalter, D., Freedman, L. et al. (1994). Bayesian approaches to clinical trials (with discussion). *Journal of Royal Statistics Society Association* **157**, 357–416.

Spiegelhalter, D., Freedman, L. et al. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials* **7**, 8–17.

Tsiatis, A., Boucher, H. et al. (1995). Sequential methods for parametric survival models. *Biometrics* **82**, 165–173.

Tsiatis, A., Rosner, G. et al. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**, 797–803.

Tsiatis, A.A., Mehta, C.R. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* **90**(2), 367–378.

URL: http://cancertrials.nci.nih.gov Policy of the National Cancer Institute for Data and Safety Monitoring of Clinical Trials.

Wald, A. (1947). *Sequential Analysis.* Wiley, New York.

Wang, S.K., Tsiatis, A.A. et al. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**(1), 193–199.

Wang, Y., Leung, D. (1997). Bias reduction via resampling for estimation following sequential tests. *Sequential Analysis* **16**, 298–340.

Wassmer, G. (2006). Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal* **48**(4), 714–729.

Wassmer, G., Vandemeulebroecke, M. (2006). A brief review on software developments for group sequential and adaptive designs. *Biometrical Journal* **48**(4), 732–737.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**, 573–581.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials.* Wiley, Chichester.

Zou, G.Y., Donner, A. et al. (2005). Group sequential methods for cluster randomization trials with binary outcomes. *Clinical Trials* **2**(6), 479–487.