Applied Bioinformatics and Biostatistics in Cancer Research

David Harrington *Editor*

Designs for Clinical Trials

Perspectives on Current Issues

Springer

Applied Bioinformatics and Biostatistics in Cancer Research

Series editors: Jeanne Kowalski, Steven Piantadosi

For further volumes:
http://www.springer.com/series/7616

David Harrington
Editor

Designs for Clinical Trials

Perspectives on Current Issues

Springer

Editor
David Harrington
Dana-Farber Cancer Institute
Boston, MA 02215
USA
david_harrington@dfci.harvard.edu

ISBN 978-1-4614-0139-1 e-ISBN 978-1-4614-0140-7
DOI 10.1007/978-1-4614-0140-7
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011936525

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to the many patients who have participated in clinical trials, with special appreciation to the selfless contribution they have made to progress in the treatment and management of chronic diseases.

Preface

In 1998, R. Smith wrote that "The randomized clinical trial is the most important advance in scientific medicine in the twentieth century." Smith's view may surprise some of our clinical colleagues but is entirely defensible. Sequences of randomized trials have led to dramatic advances in chronic diseases such as breast cancer, childhood leukemia, HIV, cardiovascular disease, and many other areas. These and other diseases are biologically complex and, despite decades of research, still understood incompletely, at best. Adequately powered trials have led to the discovery of incremental advances that have cumulatively yielded substantial improvements in survival and disease control.

The contributions to this volume examine some issues in clinical trial design that have received attention in recent years. The collection is intended primarily for the statistical practitioner, especially the very busy trialist whose active collaboration in cancer, HIV, cardiology, or other areas of medical research makes it difficult to find the time to stay abreast of advances in trial methodology. The contributors and I share the opinion that while the large majority of trials are well-designed and conducted, newer methods for design and analysis linger too long in the methodology literature before appearing in mainstream use.

The principles of design used across the range of trials are necessarily diverse; the goals of a phase I safety study are very different from a careful assessment of patient self-reported quality of life in late phase randomized trials. The common theme of the pieces in this monograph is more general rather than a specific set of tools for establishing the operating characteristics of a design. Taken together, these pieces all examine ways to make trials more efficient or more reliable and consequently more likely to reach their ultimate goal: a statistically sound contribution to evidence-based medicine that will influence practice.

The order of presentation is roughly based on the sequence of trials from early phase testing to measuring patient self-report in later phase trials. Ying-Kuen Cheung explores alternative designs for phase I trials. The methodology here is especially difficult; these trials must be small, minimize risk, and point the way

to larger efficacy trials, a seemingly impossible, or a least incompatible, set of goals. Phase I trials in cancer have stubbornly clung to the appealingly simple 3+3 designs. Cheung's work suggests that more flexibility is possible. Vince Carey and Robert Gentleman discuss the methods of randomization commonly used in trials. The randomization in a phase III trial supports causal inference about a treatment effect, so randomization is arguably the most important topic in trial design. In addition to their clear account of the principles behind many allocation schemes, Carey and Gentleman discuss their R package *randPack* implementing many randomization methods. The package can be used as a core element of a sophisticated randomization/registration system linked to a "backend" database, or by itself to study the consequences of different allocation schemes.

Sequential designs for phase III trials have matured quickly during the last 20 years, and often carry the adjective "classical" because of their widespread acceptance and use. Kyungmann Kim outlines the theory supporting both important rules of thumb (the sample size inflation factor when adding interim looks) and formal approaches (so-called α-spending functions). Institutional Review Boards or Ethics Committees almost always ask how a phase III trial will be monitored, and it is important that a trial statistician understand the basis of these designs. Two chapters then look at aspects of interim monitoring that are less well-developed, but no less important. Separately, the chapters explore two pressing questions: how might one extend a promising trial that, if sufficiently large, might detect an important treatment difference not anticipated in the initial design; and what are the consequences of stopping a trial before full information because it seems destined to be negative. Cyrus Mehta discusses the first of these topics in his chapter on methods for re-estimating sample size without increasing the type I error probability for the trial. These methods show promise in settings where abandoning a trial with substantial initial investment may be a considerable scientific and financial error. Of course, trials should not be continued when a negative outcome is very likely. Jay Herson, Marc Buyse and Janet Wittes draw on their considerable collective experience to provide guidance on early stopping for futility, outlining the methods that are available and the settings to which each is best suited. Importantly, they use case studies to illustrate the incorporation of futility analysis plans in designs.

While systemic treatments in cancer have led to many improvements in outcome, they also are accompanied by side effects and often, disappointingly, later recurrence. Rapid changes in technology have made possible measurements on the underlying molecular biology of cancer and other diseases. This increased understanding of the cellular pathways that either inhibit or accelerate proliferation of malignant cells has led to some exciting targeted therapies in cancer, most notably imatinib in chronic myelogenous leukemia and gastrointestinal stromal tumors, and trastuzumab in subtypes of breast cancer. Mei-Ling Lee reviews the currently known gene signatures that show promise either for identifying the risk of recurrence and death from particular types of cancer, or for guiding research into new therapeutics. She also explores the methodology behind the studies that validated these signatures. Stephen George and Xiaofei Wang discuss the increasingly important issue of the design of clinical trials of drugs that target subpopulations of cancer patients.

Typically, these subpopulations have been identified through a genetic or other biomarker of the type discussed by Lee, one that is either predictive or prognostic and may be a target for a therapy designed with the candidate marker in mind. These advances in molecular biology present both exciting possibilities for further progress in diseases such as cancer as well as the conundrum of ever more refined (and consequently, smaller) populations for trials. The area that George and Wang explore will likely be the subject of considerable future research.

Finally, Diane Fairclough outlines current thinking in the design and analysis of patient self-report, typically labeled but not limited to so-called quality of life measurements. Quality of life outcomes are often omitted in trial design, not because they are unimportant, but rather because they present so many challenges. The foremost of these challenges are the attrition caused by early treatment failures or general disease burden, the substantial variability in these measurements, and the complex longitudinal models that make pre-trial sample size calculations difficult. This last chapter explores all of these issues.

Clinical trialists are reminded almost daily that the participants (subjects, to use the scientific term) in these experiments often live with a potentially fatal disease and yet have agreed to help add to the understanding of treatments as partners in medical science. These participants are owed trials that converge to safe doses, finish as soon as possible, do not overlook potential advances or fail to collect data on personal responses to therapy. This book is dedicated to those participants.

Boston, MA, USA David Harrington

Reference

Smith R (1998) Fifty years of randomised controlled trials. Brit Med J 317:1166

Contents

Contributors

Marc Buyse IDDI Consultants, International Drug Development Institute, Louvain-la-Neuve, Belgium

Department of Biostatistics, Hasselt University, Diepenbeek, Belgium, marc.buyse@iddi.com

Vincent J. Carey Channing Laboratory, Harvard Medical School, Brigham and Women's Hospital, 181 Longwood Avenue, Boston, MA 02115, USA, stvjc@channing.harvard.edu

Ying Kuen Cheung Department of Biostatistics, Columbia University, 722 West 168th Street, Room 641, New York, NY 10032, USA, yc632@columbia.edu

Diane L. Fairclough Department of Biostatistics and Informatics, Colorado School of Public Health and Colorado Health Outcomes Program (COHO), School of Medicine, University of Colorado Denver, Diane.Fairclough@ucdenver.edu

Robert Gentleman Bioinformatics and Computational Biology, Genentech, Inc., 1 DNA Way South, San Francisco, CA 94080, USA, rgentlem@gene.com

Stephen L. George Duke University, Duke Box 2717, Durham, NC 27710, USA, georg001@mc.duke.edu

Jay Herson Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA, jay.herson@earthlink.net

KyungMann Kim University of Wisconsin-Madison, 600 Highland Ave, K6/438 CSC, Madison, WI 53792-4675, USA, kmkim@biostat.wisc.edu

Mei-Ling Ting Lee University of Maryland, College Park, MD 20742, USA, mltlee@umd.edu

Cyrus R. Mehta Cytel Corporation and Harvard School of Public Health, 675 Massachusetts Avenue, Cambridge, MA 02139, USA, mehta@cytel.com

Xiaofei Wang Duke University, Duke Box 2721, Durham, NC 27710, USA, xiaofei.wang@duke.edu

Janet Turk Wittes Statistics Collaborative, Inc., Washington, DC 20036, USA, janet@statcollab.com

Chapter 1
Designs for Phase I Trials

Ying Kuen Cheung

1 Introduction

The primary objective of a phase I trial of a chemotherapy is to recommend a dose for efficacy testing in phase II trials. In most phase I trials, antitumor activities will be observed but will not be a formal part of the dose-finding process. Rather, the phase II dose is often determined as the maximum dose that has an acceptable level of toxicity, the so-called maximum tolerated dose (MTD). The rationale behind this approach is that toxicities may serve as surrogate for tumor shrinkage in a patient undergoing cytotoxic cancer treatment (Ratain et al. 1993). While phase I designs for cancer trials were discussed as early as the 1960s (Carbone et al. 1965; Schneiderman 1965), formal statistical formulation of the MTD did not appear in the literature until Storer and DeMets (1987). Traditionally, a 3+3 algorithm is used for dose escalation during a trial. The algorithm starts at a low dose (e.g., one-tenth of LD10 in mice) and enrolls patients in groups of three: the next group of three will receive the next higher dose if no toxicity occurs in the current group, and stay at the same dose if one patient experiences toxicity; the trial will stop if there are two or more patients experiencing toxicity at a dose. The MTD is identified by the largest test dose with fewer than two patients experiencing a dose-limiting toxicity during the first course of chemotherapy. There is no clear justification for using the 3+3 method apart from convention. A main criticism is that its operating characteristics depend arbitrarily on the underlying dose-toxicity curve and the MTD thus identified does not correspond to any interpretable quantity upon repeated sampling. From an ethical viewpoint, the conservative 3+3 escalation scheme tends to place many patients at low and hence inefficacious doses, and may cause difficulty

Y.K. Cheung (✉)
Department of Biostatistics, Columbia University, 722 West 168th Street, Room 641, New York, NY 10032, USA
e-mail: yc632@columbia.edu

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7_1,

in consenting patients early in the trial (Rosa et al. 2006). Since the late 1980s, several dose-finding designs, primarily motivated by applications to oncology, have been proposed to address these problems of the 3+3 method (Babb et al. 1998; O'Quigley et al. 1990; Simon et al. 1997; Storer 1989). The recent literature has seen proposals of novel phase I trial designs for other disease areas such as acute stroke (Cheung 2007; Cheung and Elkind 2010; Elkind et al. 2008), amyotrophic lateral sclerosis (Cheung et al. 2006), and HIV (O'Quigley et al. 2001). While toxicity does not play the same central role in drugs for other diseases as in cytotoxic agents, the MTD still serves as a useful safety limit for future clinical investigation. In this chapter, we provide a systematic review of dose-finding methods for the "standard" phase I trials where the primary objective is to determine the MTD defined in terms of binary toxicities over a short-term period.

Broadly, phase I designs can be classified by the dose-escalation approaches as algorithm-based (Sect. 2) or model-based (Sect. 3). An algorithm-based design prescribes a set of dose-escalation rules for any given dose without regard to the observations at other doses. It has the virtue of simplicity because the rules can be tabulated and made accessible to the clinical investigators prior to the start of a trial. In contrast, a model-based design makes explicit use of a dose-toxicity model in the dose escalation decisions, thus allowing borrowing strength from information across doses. In Sect. 4, we compare some promising designs by simulations in the context of a chemotherapy trial in lymphoma patients (Leonard et al. 2005). Section 5 presents some key theoretical criteria for these dose-finding designs.

Another possible taxonomy for phase I designs is based on the study endpoints chosen in accordance with the clinical situations such as the nature of toxicity, the treatment plan, and the drug mechanism. In Sect. 6, we will review designs for three types of "nonstandard" endpoints. First, we will discuss and advocate the use of time-to-toxicity endpoints to deal with situations with delayed toxicity, which is ubiquitous in multi-course chemotherapy, radiation therapy, and chemoprevention. Second, we will review methods that use the ordinal toxicity severity weights, as opposed to the dichotomized toxicity outcomes. Differentiating the severity among high grade toxicities is crucial when giving high-risk experimental regimens such as high-dose chemotherapy that may cause irreversible and disabling side effects or even death, as well as severe but reversible toxicity. Third, we will outline the basic concepts of bivariate designs that use both toxicity and efficacy endpoints for situations where toxicity-driven designs may not adequately answer relevant dosing questions. To make these novel designs accessible to the clinical colleagues, statisticians need to be able to deliver and explain the methods in practice. Section 7 will discuss guidelines for method implementation and present the challenge ahead.

2 Algorithm-Based Designs

A prominent example of algorithm-based designs is the 3+3 algorithm. Despite its deficiencies, the method is by far the most widely used design. It is common to treat additional patients (typically 6–12) at the identified MTD to improve

estimation precision for preliminary efficacy. Given the method's poor dose selection properties, the additional patients will likely be treated at an inappropriate dose. The fundamental problem is the 3+3 method lacks a quantitative objective. This section reviews some recent algorithm-based designs with rigorously defined objectives.

2.1 Sequential Stepwise Tests

Consider a trial with K doses, and let p_i denote the probability of toxicity at dose i. The MTD is defined as the largest dose with $p_i \leq \theta$ for some target θ. The design objective is to identify the MTD with a high probability and keep the probability of selecting an unsafe dose low. Precisely, dose i is said to be safe if $p_i < \phi$ for some $\phi > \theta$. Then, the maximum safe dose (MAXSD) is defined as $\nu = \max\{i : p_i < \phi\}$ with the convention $\max\emptyset = 0$, and can be estimated by a stepwise test $\hat{\nu}$ with respect to the family of hypotheses: $H_{0i} : p_i \geq \phi$ versus $H_{1i} : p_i < \phi$. Under this formulation, an unsafe dose will be declared safe and selected when a type I error occurs against any true H_{0i}. Therefore, the probability of selecting an unsafe dose can be equivalently controlled via the familywise error rate (FWER) of $\hat{\nu}$ defined as

$$\text{FWER}(\hat{\nu}) = \Pr(\hat{\nu} > \nu) \equiv \max_{0 \leq m \leq K} \sup_{\mu \in \Theta_m} P_\mu(\hat{\nu} > m),$$

where P_μ is the probability computed under the toxicity vector $\mu = (p_1, \ldots, p_K)^T$ and $\Theta_m = \{\mu : p_m < \phi, p_k \geq \phi; k > m\}$ is the parameter subspace in which $\nu = m$. In addition, the probability of correct selection of the procedure $\hat{\nu}$ is defined as

$$\text{PCS}(\hat{\nu}) = \min_{0 \leq m \leq K} \inf_{\mu \in \Theta_m^*} P_\mu(\hat{\nu} = m),$$

where $\Theta_m^* = \{\mu : \max(p_1, \ldots p_m) \leq \theta, p_k \geq \phi; k > m\}$ is a subset of Θ_m under which the MTD is equal to ν.

Using this multiple testing framework, Cheung (2007) proposes two-stage stepwise procedures that allow sequential dose assignments as ethics requires. The first stage starts at dose $i = S_1 < K$ and escalates to $i + 1$ if and only if the hypothesis H_{0i} is rejected; we will use $\mathscr{R}_{i1}(\mathscr{A}_{i1})$ to denote the rejection (acceptance) region of H_{0i} by data collected at dose i at stage 1. Escalation stops once $\mathscr{A}_{i1}$ is observed for some i. The second stage starts at $S_2 = \min\{i : \mathscr{A}_{i1} \text{ is observed}\} - 1$ with additional patients; and deescalation occurs until $\mathscr{R}_{i2}$ is observed for some i. Here, $\mathscr{R}_{i2}$ denotes the rejection region of H_{0i} based on the cumulative data collected at dose i at the end of stage 2. The MTD is estimated by $\hat{\nu} = \max\{i : \mathscr{R}_{i1} \cap \mathscr{R}_{i2} \text{ is observed}\}$.

For any given $\mathscr{R}_{il}$, the distribution of $\hat{\nu}$ is computed as

$$P_\mu(\hat{\nu} = m) = \begin{cases} \delta_{m2} \left[\prod_{j=m+1}^{S_1-1}(1-\delta_{j2})\right] \Omega_{S_1,K} & \text{for } 0 \leq m \leq S_1 - 1 \\ \left[\prod_{j=S_1}^{m-1} \delta_{j1}\right] \gamma_m \Omega_{m+1,K} & \text{for } S_1 \leq m \leq K \end{cases},$$

where $\delta_{il} = \Pr(\mathscr{R}_{il}|p_i)$ for $l = 1,2$, $\gamma_i = \Pr(\mathscr{R}_{i1} \cap \mathscr{R}_{i2}|p_i)$, $\delta_{02} \equiv 1, \Omega_{K+1,K} \equiv 1$, and

$$\Omega_{m,K} = (1-\delta_{m1}) + \sum_{i=m}^{K-1} (1-\delta_{i+1,1}) \prod_{j=m}^{i} (\delta_{j1} - \gamma_j) + \prod_{j=m}^{K} (\delta_{j1} - \gamma_j).$$

Furthermore, operating characteristics of $\hat{\nu}$ can be summarized and computed as $\text{FWER}(\hat{\nu}) = 1 - P_{\mu_0^*}(\hat{\nu} = 0)$ and $\text{PCS}(\hat{\nu}) = \min_{0 \le m \le K} P_{\mu_m^*}(\hat{\nu} = m)$, where μ_m^* is the toxicity vector with $p_i = \theta$ for $i \le m$ and $= \phi$ for $i > m$, provided that $\mathscr{R}_{i1}$ and $\mathscr{R}_{i1} \cap \mathscr{R}_{i2}$ are unbiased tests H_{0i} against H_{1i}, that is, δ_{i1} and γ_i are decreasing in p_i.

We can construct unbiased test regions using the likelihood ratio test (LRT) and let $\mathscr{R}_{il} = \{Z_i(N_l) \le c_l\}$ for some prespecified $c_1 \le c_2$ and $N_1 \le N_2$, where $Z_i(j)$ denotes the number of toxic outcomes in the first j patients at dose i. The stepwise LRT procedure is defined completely by the parameters (c_1, N_1, c_2, N_2). It is quite straightforward to iterate and find the set of design parameters that give the smallest expected number of patients enrolled to a dose to reach a conclusion about H_{0i} under $p_i = \phi$, denoted by $E_\phi(N^*)$, among all LRT that satisfy

$$\Pr(\mathscr{R}_{i1}|\phi) \le \varepsilon^*, \text{FWER}(\hat{\nu}) \le \alpha_0 \text{ and } \text{PCS}(\hat{\nu}) \ge 1 - \alpha_1. \tag{1.1}$$

For LRT, $E_\phi(N^*) = N_1 \Pr(\mathscr{A}_{i1}|\phi) + N_2 \Pr(\mathscr{R}_{i1}|\phi)$. Generally, $E_\phi(N^*)$ is proportional to the expected total number of patients receiving an overdose in a trial. A stepwise procedure that minimizes $E_\phi(N^*)$ is referred to as a minimum overdose design.

Alternatively, the sequential probability ratio test (SPRT) (Wald 1945) prescribes the test regions $\mathscr{R}_{il} = \{\lambda_{i,\tau_{il}} \ge \rho_l\}$ and $\mathscr{A}_{il} = \{\lambda_{i,\tau_{il}} \le \xi\}$ for $l = 1,2$, where

$$\lambda_{i,n} = \frac{\theta^{Z_i(n)}(1-\theta)^{n-Z_i(n)}}{\phi^{Z_i(n)}(1-\phi)^{n-Z_i(n)}}$$

is the likelihood ratio for dose i, and $\tau_{il} = \inf\{n > 0 : \lambda_{i,n} \ge \rho_l \text{ or } \lambda_{i,n} \le \xi\}$. Cheung (2007) provides an algorithm to find (ρ_1, ρ_2, ξ) that minimizes $E_\phi(N^*)$ subject to the constraints (1.1), and obtain the minimum overdose SPRT design.

Consider a trial starting at $S_1 = 3$ among $K = 5$ doses. With $\alpha_0 = 0.25$, $\alpha_1 = 0.50$, $\varepsilon^* = 0.50$, $\theta = 0.25$, and $\phi = 0.45$, the LRT attains a minimum $E_\phi(N^*) = 13.2$ when $c_1 = 3, N_1 = 8, c_2 = 5$, and $N_2 = 19$, and the SPRT attains a minimum $E_\phi(N^*) = 9.5$ when $\rho_1 = 1.683$, $\rho_2 = 9.325$, and $\xi = 0.317$. In this example, the SPRT has a smaller $E_\phi(N^*)$ than the LRT as expected because the SPRT is known to be optimal under the hypothesized values (Wald and Wolfowitz 1948). Extensive simulations verify that the stepwise SPRT is generally superior to the LRT in terms of average sample size and PCS. However, since the SPRT is an open-ended procedure, truncated test should be applied for practicality. The maximum sample size (N_2) in the minimum overdose LRT provides a justified truncation for the SPRT.

A potential limitation of the stepwise tests seems to be a large maximum sample size (KN_2). In the above designs, a maximum of 19 patients per dose is much greater

Table 1.1 Dose decisions at each sample size n_k by the minimum overdose SPRT design for a trial starting at $S_1 = 3$ among $K = 5$ doses, with $\alpha_0 = 0.25, \alpha_1 = 0.50, \varepsilon^* = 0.50, \theta = 0.25$, and $\phi = 0.45$. The SPRT is truncated at 19 subjects

Stage 1: Intervals of Z_k				Stage 2: Intervals of Z_k			
n_k	Escalate	Stay	Deescalate	n_k	Stop[a]	Stay	Deescalate
2	0	1	2	7	–	[0,3]	[4,7]
3	0	[1,2]	3	8	0	[1,4]	[5,8]
4	0	[1,3]	4	9	0	[1,4]	[5,9]
5	[0,1]	[2,3]	[4,5]	10	0	[1,4]	[5,10]
6	[0,1]	[2,3]	[4,6]	11	[0,1]	[2,5]	[6,11]
7	[0,1]	[2,3]	[4,7]	12	[0,1]	[2,5]	[6,12]
8	[0,2]	[3,4]	[5,8]	13	[0,2]	[3,5]	[6,13]
9	[0,2]	[3,4]	[5,9]	14	[0,2]	[3,6]	[7,14]
10	[0,2]	[3,4]	[5,10]	15	[0,2]	[3,6]	[7,15]
11	[0,3]	[4,5]	[6,11]	16	[0,3]	[4,6]	[7,16]
12	[0,3]	[4,5]	[6,12]	17	[0,3]	[4,7]	[8,17]
...	...	...	...	18	[0,3]	[4,7]	[8,18]
19[b]	[0,5]	–	[6,19]	19[b]	[0,5]	–	[6,19]

[a]Once the stopping criteria is reached, the dose is confirmed safe
[b]The boundaries are modified for truncation

than what a 3+3 trial will require (maximum 6 per dose), although simulations show that the stepwise SPRT often concludes a trial with reasonable sample sizes. While a maximum sample size that is larger than what is perceived to be feasible may turn investigators away from using the method, the sample size determination reflects the inadequacy of the sample size in current practice. The stepwise test is the first, and so far the only, phase I design for which the sample size can be justified analytically with respect to frequentist properties without depending entirely on simulations. The succinct error statements (1.1) help clinicians appreciate the statistical inputs in a dose-finding study, and facilitate a defensible approach to design the study. Also, the sequential procedure is easy to implement and operates in a manner similar to the standard algorithm. The dose decisions by the stepwise test can be charted and made available to clinicians prior to a trial; see Table 1.1 for the SPRT procedure.

2.2 Up-and-Down Designs via Random Walk

Storer (1989) describes some up-and-down dose-escalation schemes, and proposes the use of combinations of schemes in stages. For example, his design BD consists of

Design B. Single patients are treated. The next patient is treated at the next lower dose level if a toxic response is observed, otherwise at the next higher dose level.

Design D. Groups of three patients are treated. Escalation occurs if no toxicity is seen and deescalation if more than one patient has toxicity. If a single patient has toxicity, the next group of three is treated at the same level.

The first stage of design BD follows design B, which is intended to move the trial quickly through the low doses. Once toxicity is seen, the trial switches to design D for dose escalation until a prespecified number of patients has been enrolled. Using Markov Chain representation, it can be shown that design BD (design D) tends to sample around a dose that causes toxicity with a probability $\theta = 0.33$. Therefore, design BD seems to be an appropriate method if the MTD is defined as the 33rd percentile. Storer further suggests fitting the toxicity data to a logistic curve after the trial ends, and estimates the MTD by maximum likelihood estimate of the 33rd percentile of the fitted curve. Since the design concentrates sampling around the target, the estimation is expected to be efficient and robust even if the logistic model is not a correct specification of the dose-toxicity curve over the whole dosing range.

Durham et al. (1997) propose a biased coin design for any target $\theta \leq 0.50$. The design will treat the next patient at the next lower dose if the current patient has toxicity, and at the next higher dose with probability $\theta/(1-\theta)$ if otherwise. This method may be viewed as a randomized extension of the Dixon and Mood's up-and-down method (Dixon and Mood 1948), and can be shown to produce a unimodel asymptotic treatment allocation around the dose closest to the target percentile; see Theorem 2 in Durham and Flournoy (1994).

These up-and-down rules assign a dose for the next patient (or group of patients) on the basis of the most current group, without regard to the previous outcomes at the same dose. As a result of this Markov property, their sampling properties is analytically tractable for any given toxicity configuration via the transition matrix. For example, for the biased coin design, the transition probabilities of deescalation, staying, and escalation from a dose $i = 2, \ldots, K-1$ are p_i, $(1-p_i)(1-2\theta)/(1-\theta)$, and $(1-p_i)\theta/(1-\theta)$, respectively; the transition probabilities from doses 1 and K can also be easily determined. However, frequentist properties such as (1.1) are yet to be examined inductively over many configurations. Also due to the Markov property of random walk, any dose, however, toxic it appears, may be revisited with a nonnegligible probability even as data is accrued throughout the trial. Such lack of convergence of treatment allocation may cause ethical difficulties in human trials.

2.3 Algorithm per Toxicity Probability Intervals

Ji et al. (JLB) (2007) consider an up-and-down design that assign doses based on the posterior intervals of toxicity probability. The authors suggest using a noninformative Beta prior with shapes $(0.005, 0.005)$ on p_i so that the posterior of p_i is also Beta with shapes $(0.005 + Z_i, 0.005 + n_i - Z_i)$, where n_i and $Z_i = Z_i(n_i)$, respectively, denote the sample size and the number of toxicities at dose i. The posterior probabilities favoring escalation (E), staying (S), and deescalation (D) from a dose i are then defined, respectively, as $q(E,i) = \Pr[p_i - \theta < -K_1\sigma_i | n_i, Z_i]$, $q(S,i) =$

Table 1.2 Dose decisions by the JLB (2007) method for $\theta = 0.25$ with $K_1 = 1.0, K_2 = 1.5$

	Intervals of Z_k for each dose decision			
n_k	Escalate	Stay	Deescalate	Deescalate and close
1	0	–	–	1
2	0	1	–	2
3	0	1	2	3
4	0	1	2	[3,4]
5	0	[1,2]	3	[4,5]
6	0	[1,2]	3	[4,6]
7	0	[1,3]	–	[4,7]
8	0	[1,3]	4	[5,8]
9	[0,1]	[2,3]	4	[5,9]
10	[0,1]	[2,4]	–	[5,10]

$\Pr[-K_1\sigma_i \le p_k - \theta \le K_2\sigma_i | n_i, Z_i]$, and $q(D,i) = \Pr[p_i - \theta > K_2\sigma_i | n_i, Z_i]$, where σ_i is the posterior standard deviation of p_i and $K_1, K_2 > 0$ are prespecified. The dose decision (E,S,D) with the highest posterior probability will be assigned to the next patient group. An exception is when $q(E,i) > \max\{q(S,i), q(D,i)\}$ but dose $i+1$ is proven unacceptably toxic with $\Pr(p_{i+1} > \theta | n_{i+1}, Z_{i+1}) > 0.95$, then the dose for the next group will stay at the current dose i, with $i+1$ permanently closed. At the end of the trial, the MTD is estimated by the dose with $E(p_i | n_i, Z_i)$ closest to θ among all acceptable doses.

In contrast to the random walk rules, JLB use all observations accrued to a dose to make a dose decision, and have provisions for closing a toxic dose so that it will not be revisited. At the same time, since the dose decision is made without regard to observations at other doses, it is possible to enumerate the prescribed action in a table that the clinical investigators may use during a trial without performing the posterior computations. For example, Table 1.2 displays the decision intervals of Z_k for a trial targeting at $\theta = 0.25$.

3 Model-Based Designs

The continual reassessment method (CRM) (O'Quigley et al. 1990) is among the first model-based phase I designs and has generated an extensive literature of its own. The basic idea is to make dose decisions based on a dose-toxicity curve that is being continually reassessed as data is accrued during a trial. This approach has drawn attention in the medical community (Ratain et al. 1993) and appeals to clinicians as an ethical alternative to the standard phase I method (Rosa et al. 2006). The subsequent proposals of model-based designs by and large adopt the notion of continual reassessment. This section thus provides a review of the model-based methods with an emphasis on the CRM.

3.1 The Continual Reassessment Method

3.1.1 The Basic Approach

In a typical phase I trial, patients are treated at a discrete panel of doses, denoted by $d_1, \ldots, d_K$. Let Y_j be the toxicity indicator of the jth patient enrolled to the trial, and x_j the dose assigned to the patient. The CRM assumes a one-parameter model $F(x, \beta)$ that is strictly increasing in dose x and monotone in the parameter β, so that $p_k = \Pr(Y_j = 1 | x_j = d_k)$ is postulated to be $F(d_k, \beta)$ for some β. Since the CRM is originally presented as a Bayesian approach, a prior distribution $G(\beta)$ is assumed. The trial starts by treating the first patient at the prior MTD ν_0, that is, $x_1 = d_{\nu_0}$, where $F(d_{\nu_0}, \hat{\beta}_0) = \theta$ and $\hat{\beta}_0 = \int \beta \mathrm{d}G(\beta)$ is the prior mean of β. With the prior and data in the first n patients, β is estimated by its posterior mean (denoted by $\hat{\beta}_n$), and the next patient is set to receive $x_{n+1} = \arg\min_{d_k} |F(d_k, \hat{\beta}_n) - \theta|$ for some prespecified target rate θ. This process continues until a desired sample size N is reached. The final MTD estimate is given by x_{N+1}.

Common choices of dose-toxicity model $F(x, \beta)$ in the CRM literature include the empiric function x^β and the logistic function $\{1 + \exp(-a_0 - \beta x)\}^{-1}$ with a fixed intercept a_0. Due to the monotone dose-toxicity assumption, the parameter β is restricted to take on positive values, and is typically assumed lognormal or gamma a priori. From a computational viewpoint, we find Gaussian quadrature (Naylor and Smith 1982) provides an accurate approximation for $\hat{\beta}_n$ in the following parametrization:

$$F(x, \beta) = x^{\exp(\beta)} \text{ (empiric) and } F(x, \beta) = \left\{1 + \exp\left(-a_0 - e^\beta x\right)\right\}^{-1} \text{ (logistic)}, \tag{1.2}$$

where β has a normal prior. This model setup also allows approximation of the maximum likelihood approach (O'Quigley and Shen 1996) by specifying a large variance in the prior.

3.1.2 Practical Modifications

The CRM is motivated by applications to phase I trials conducted in cancer patients. To avoid treating many patients at low and inefficacious doses, the method starts at a dose believed to be close to the true MTD. As a result, the original CRM may cause more toxic outcomes than the standard 3+3 design. Expectedly, it raises safety concerns and draws criticisms (Korn et al. 1994). Modifications have been suggested to enhance the safety and practicality of the CRM. A common modification is the adoption of an initial stage that starts at a low dose until the first toxicity occurs. Generally, an initial design can be specified by a predetermined nondecreasing sequence $\{x_{n,0}\}$ with $x_{j-1,0} \leq x_{j,0}$. A two-stage CRM assigns to the next patient

$$x_{n+1} = \begin{cases} x_{n+1,0} & \text{if } Y_j = 0 \text{ for all } j \le n, \\ \arg\min_{d_k} |F(d_k, \hat{\beta}_n) - \theta| & \text{if } Y_j = 1 \text{ for some } j \le n. \end{cases}$$

For example, a two-stage design with $x_{n,0} = \lceil n/3 \rceil$ for $n = 1, \ldots, 3K$ and $= K$ for $n = 3K+1, \ldots, N$ initially escalates dose after every three nontoxic outcomes, and switches to the CRM for dose decision once the first toxicity is seen; where $\lceil x \rceil$ is the smallest integer that is larger than or equal to x.

Another safety restriction is often imposed to prevent dose escalation for the next patient if the current patient experiences a toxic outcome. This restriction ensures dose-finding coherence which will be further discussed in Sect. 5.1.

3.1.3 Model Specification

An important practical point about the CRM is that the doses $d_1, \ldots, d_K$ are not the actual doses administered, but are defined on a conceptual scale that represents an ordering of the risks for toxicity. This conceptual scale is useful, for instance, in combination trials where each subsequently higher dose involves incrementing doses of different treatments and there is no natural scale of dosage. In practice, d_k is obtained by substituting the initial guess of toxicity probability p_{0k} for dose level k into the dose-toxicity model, that is, $p_{0k} = F(d_k, \hat{\beta}_0)$. Consider a trial with $K = 5$ doses and a target $\theta = 0.25$. Suppose, we use the empiric function in (1.2) as the working dose-toxicity model with $\beta \sim N(0, 1.34)$ a priori, and believe that $\nu_0 = 3$ is the prior MTD with $p_{03} = 0.25$. Then, d_3 is determined such that

$$d_3^{\exp(0)} = 0.25.$$

The remaining d_i's can be determined in a similar manner in accordance with the corresponding initial guesses p_{0i}'s. Thus, a CRM model includes the specification of $p_{01}, \ldots, p_{0K}$ in addition to the dose-toxicity function F and prior G.

Ideally, the initial guesses are specified to reflect the clinician's prior beliefs. In practice, this is hardly achieved and p_{0k}'s are chosen as design parameters that yield good operating characteristics. Intuitively, the CRM will likely recommend the true MTD ν if the true dose-toxicity curve is steep around ν. Let β_k be defined such that $F(d_k, \beta_k) = p_k$. Then, the "steepness condition" is mathematically represented by

$$\beta_\nu \in (b_\nu, b_{\nu+1}), \beta_k \in \bigcup_{i=k+1}^{K} (b_i, b_{i+1}) \text{ for } k < \nu, \text{ and } \beta_k \in \bigcup_{i=1}^{k-1} (b_i, b_{i+1}) \text{ for } k > \nu, \tag{1.3}$$

where b_k solves $F(d_{k-1}, b_k) + F(d_k, b_k) = 2\theta$ for $k = 2, \ldots, K$, and b_1 and b_{K+1} are the limits the parameter space of β. Consider the empiric function with $p_{01} = 0.05$, $p_{02} = 0.12$, $p_{03} = 0.25$, $p_{04} = 0.40$ and $p_{05} = 0.55$, and $\theta = 0.25$. Suppose $\nu = 3$ and confine β to a sufficiently wide but finite interval, say $[-5, 5]$.

Table 1.3 Indifference intervals for four CRM models for $\theta = 0.25$ with $K = 5$. Models E1 and L1 are, respectively, the empiric function and logistic function defined in (1.2) with initial guesses (0.05,0.12,0.25,0.40,0.55), and models E2 and L2 with initial guesses (0.05,0.10,0.25,0.45,0.80). Models E1 and L1, having narrower indifference intervals, are more sensitive than E2 and L2

ν	Model E1	Model E2	Model L1	Model L2
1	(–,0.31)	(–,0.30)	(–,0.32)	(–,0.30)
2	(0.19,0.32)	(0.21,0.34)	(0.18,0.33)	(0.20,0.34)
3	(0.18,0.32)	(0.16,0.34)	(0.18,0.32)	(0.16,0.34)
4	(0.18,0.32)	(0.16,0.45)	(0.18,0.32)	(0.16,0.46)
5	(0.18,–)	(0.06,–)	(0.18,–)	(0.04,–)

Then, condition (1.3) becomes $\beta_1 \in (-0.59, 5)$, $\beta_2 \in (-0.20, 5)$, $\beta_3 \in (-0.20, 0.22)$, $\beta_4 \in (-5, 0.22)$, and $\beta_5 \in (-5, 0.63)$. Converting these intervals in the parameter space for β to intervals on the probability scale gives $p_1 \in (0, 0.19)$, $p_2 \in (0, 0.18)$, $p_3 \in (0.18, 0.32)$, $p_4 \in (0.32, 1)$, and $p_5 \in (0.33, 1)$. Condition (1.3) is violated if, for instance, $p_2 \in (0.18, p_3)$ with $|p_3 - 0.25| < |p_2 - 0.25|$. In this case, the CRM may select dose 2 instead of $\nu = 3$. Choosing a dose with $p_2 \in (0.18, p_3)$ may be an erroneous selection but not a serious one. The same can be said if $p_4 \in (p_3, 0.32)$. The interval (0.18, 0.32) is called the indifference interval for $\nu = 3$ because the CRM model may fail to differentiate ν and its neighbors with toxicity probabilities falling in this interval.

Different CRM models have different indifference intervals, and apparently, models with large indifference intervals are not sensitive enough to detect relevant difference in toxicity probabilities. Table 1.3 shows the indifference intervals for four CRM models with $\theta = 0.25$ using different functional forms and different sets of initial guesses. We note that the impacts of the functional form on sensitivity are small in comparison to that of the initial guesses. Simulation is the primary tool to evaluate the operating characteristics of a CRM design. With infinite possibilities of choices of the initial guesses, it is thus sensible to restrict consideration to those with adequate sensitivity. As such, the indifference interval technique provides a quick supplement to simulation when planning a CRM trial.

3.2 Escalation with Overdose Control

Other model-based designs by and large use the same idea that continually updates the belief about the dose-toxicity relationship during a trial. The main difference lies in the dose decision criterion. Babb et al. (1998) propose the escalation with overdose control (EWOC) in which the next dose x_{n+1} is given so that the posterior probability of x_{n+1} exceeding the MTD is equal to α. The method assumes a dose-toxicity model $F(x, \beta)$ where β is vector-valued (e.g., two-parameter logistic) so that the model is more flexible and "realistic" than that in the CRM. Then the model-based MTD $x_{\theta,\beta}$ is defined such that $F(x_{\theta,\beta}, \beta) = \theta$. Under a Bayesian

paradigm, $x_{\theta,\beta}$ is random via the postulation of prior on the model parameter β, and x_{n+1} is defined such that $Q_n(x_{\theta,\beta} \le x_{n+1}) = \alpha$ with $Q_n(\cdot)$ denotes probability computed under the posterior distribution of β given the first n observations. From a decision-theoretic viewpoint, x_{n+1} minimizes the Bayes risk with respect to the loss function

$$L_\alpha(x_\theta, x) = \begin{cases} \alpha(x_\theta - x) & \text{if } x \le x_\theta, \text{ i.e., if } x \text{ is an underdose} \\ (1-\alpha)(x - x_\theta) & \text{if } x > x_\theta, \text{ i.e., if } x \text{ is an overdose} \end{cases},$$

where x is an estimate of the true x_θ. This loss function implies that the loss incurred by treating a patient at one unit above the MTD is $(1-\alpha)/\alpha$ times greater than that by treating a patient at one unit below the MTD. Thus, a small α (e.g., 0.05) puts a heavier penalty on overdose than underdose, and approaches the target dose from below. From a practical viewpoint, the applicability of EWOC may be limited to situations where a continuum of doses are available (e.g., IV administration).

3.3 Curve-Free CRM

In an attempt to avoid the reliance of a parametric dose-toxicity model, Gasparini and Eisele (2000) propose a phase I design that governs dose assignments during a trial by Bayesian nonparametric estimates of the dose-toxicity curve. A product-of-beta prior is put on the toxicity probabilities $p_1, \ldots, p_K$, namely,

$$1 - p_1, \frac{1-p_2}{1-p_1}, \frac{1-p_3}{1-p_2}, \ldots, \frac{1-p_K}{1-p_{K-1}}$$

are assumed independently distributed as beta variables a priori. Besides monotone dose-toxicity imposed by this prior, there is no parametric assumption on the toxicity probabilities. The beliefs on p_i's are reassessed after every small group of patients so that the next group is treated at $\arg\min_i |E(p_i|\text{ data}) - \theta|$. This method proceeds in the same manner as the CRM, and is usually called the curve-free CRM (CFM).

The CFM may cause rigidity that confines dose assignments to suboptimal levels (Cheung 2002). Specifically, such rigidity occurs when a dose i is observed with large observed toxicity rates (e.g., one toxic outcome out of two patients): a high estimate for p_i will drive the sequential CFM to assign a lower dose (dose $i-1$, say), and the nonparametric estimation will prevent escalation back to dose i because whatever happens at dose $i-1$ will have little effect on the estimation of p_i. Such rigidity may occur with a nonnegligible probability even if dose i is in truth safe (i.e., $p_i \le \theta$). This problem can be alleviated by using informative priors on p_i's and larger group size (Cheung 2002). As a special case, the CFM reduces to the CRM when using degenerated priors (whichessentially impose a parametric

structure on p_i's). In this light, the use of nonparametric estimation in conjunction with a CRM-type sequential design is not recommended. Rather, the CRM with a one-parameter model is a reasonable method of choice, when calibrated using the sensitivity technique in Sect. 3.1.3.

4 Example: A Bortezomib Trial in Lymphoma Patients

We designed a dose-finding study of bortezomib when used in combination with the standard chemotherapy regimen as first-line treatment for lymphoma patients (Leonard et al. 2005). Dose-limiting adverse events were defined as grade 3 or more severe neuropathy, low platelet count, and symptomatic nonneurologic or nonhematologic toxicity. The objective of the trial was to identify the MTD among $K = 5$ test dose schedules with a target $\theta = 0.25$; each subsequent regimen increases in intensity by escalating the dose of bortezomib or the administration frequency. The trial started the first subject at the third level ($0.7\,\mathrm{mg\,m^{-2}}$ days 1 and 4 of a 21-day cycle), and a modified CRM (Sect. 6.1) was used for subsequent dose escalation in 18 subjects. The empiric model in (1.2) was used with $\beta \sim N(0, 1.34)$; see model E1 in Table 1.3.

4.1 Method Comparison

To anticipate how the CRM operates, Fig. 1.1a shows the outcomes of a simulated trial run under the scenario $p_1 = p_2 = 0.05$, $p_3 = 0.08$, $p_4 = 0.25$, and $p_5 = 0.45$. In the trial, each patient was enrolled with a latent tolerance uniformly distributed on the interval $(0, 1)$. If the uniform variate was smaller than the toxicity probability associated with the dose given to the patient, then the patient had a toxic outcome; otherwise, the patient did not have a toxic outcome. The trial started at dose level 3, and dose assignments in the early part of the trial went back and forth between dose level 4 and its neighboring doses. After patient 8, the CRM treated the remaining patients at dose 4, which was also the final estimate of the MTD.

Figure 1.1b illustrates how the 3+3 algorithm might draw a wrong conclusion. To match it with the simulated CRM trial, the same sequence of uniform variates was applied in the generation of toxicity outcomes. Also for comparison purposes, the algorithm was started at dose level 3. The trial ended after observing two toxic outcomes at level 4, with a total of 6 enrolled patients. Simple binomial calculation shows that, with $p_4 = 0.25$, there is a nonnegligible probability (0.14) to observe two toxic outcomes out of three subjects. While this is the sample size clinicians expect to see, the 3+3 algorithm in this case concludes dose 4 as toxic too soon.

While the CRM is expected to outperform the 3+3 algorithm, it is of interest to consider some recent algorithm-based designs whichare logistically simpler than

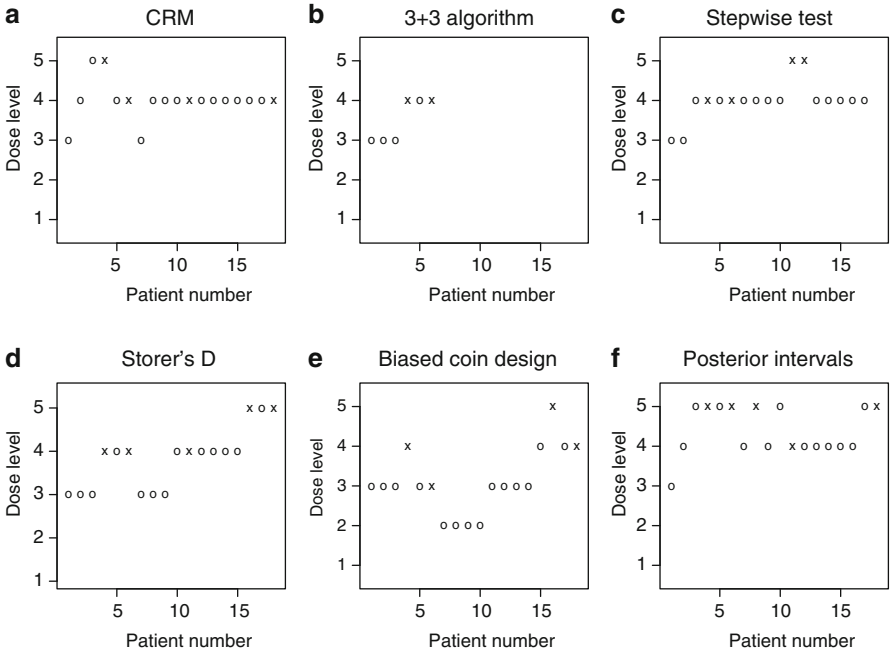

Fig. 1.1 Simulated trials by six dose-escalation methods. Each point represents a patient, with "o" indicating no toxicity and "x" indicating toxicity. Except for the 3+3 algorithm and the stepwise procedure, the sample size is prespecified to be 18

the model-based CRM. Figure 1.1c shows the outcomes for the stepwise SPRT as described in Table 1.1, Fig. 1.1d for Storer's design D, Fig. 1.1e for the biased coin design, and Fig. 1.1f for the JLB method as in Table 1.2. All designs were started at dose 3, and the same sequence of uniform variates was used to generate the toxicity outcomes. The biased coin design exhibited much within-trial variation of dose assignments, primarily due to the Markovian nature of the design: after the fifth patient at dose 3 experiencing a toxic outcome, the next dose was deescalated, despite the four previous nontoxic observations at dose 3. The same "memoryless" property may cause Storer's design D to deescalate when the current dose appears safe, or escalate when it appears toxic, perhaps to a lesser extent because of its group-accrual nature. The stepwise SPRT induced less within-trial dose variation than the up-and-down designs, while it allowed quick escalation after two nontoxic outcomes at dose 3, and deescalation after two toxic outcomes at dose 5. The SPRT trial ended with 13 subjects enrolled to dose 4. The JLB method exhibited more within-trial variation than the stepwise procedure, but less when compared to the memoryless designs. It assigned 9 patients to dose 4. Both the stepwise test and JLB method selected dose 4 as the MTD.

Table 1.4 Operating characteristics of the CRM, the stepwise SPRT, the JLB and the 3+3 methods

	Proportion of selecting dose					Average numbers	
Design	0/1[a]	2	3	4	5	Toxicity	Sample size
Scenario 1, **p**:	0.05	0.05	**0.25**	0.45	0.55		
3+3	0.05	0.39	**0.43**	0.12	0.02	2.7	14
CRM; $n = 18$	0.01	0.17	**0.65**	0.17	0.01	4.7	18
JLB; $n = 18$	0.03	0.28	**0.45**	0.20	0.03	4.4	18
CRM; $n = 30$	0.00	0.10	**0.77**	0.13	0.00	7.7	30
JLB; $n = 30$	0.04	0.29	**0.53**	0.13	0.00	6.8	30
SPRT; $n_{tr} = 19$	0.00	0.22	**0.71**	0.07	0.00	10	32
Scenario 2, **p**:	0.05	0.05	0.08	**0.25**	0.45		
3+3	0.05	0.06	0.35	**0.41**	0.12	2.6	16
CRM; $n = 18$	0.00	0.01	0.22	**0.61**	0.16	4.2	18
JLB; $n = 18$	0.01	0.07	0.26	**0.50**	0.16	3.9	18
CRM; $n = 30$	0.00	0.01	0.15	**0.73**	0.12	7.1	30
JLB; $n = 30$	0.02	0.07	0.29	**0.54**	0.08	6.2	30
SPRT; $n_{tr} = 19$	0.00	0.01	0.22	**0.71**	0.07	9.0	31
Scenario 3, **p**:	0.05	0.05	0.08	0.12	**0.25**		
3+3	0.05	0.07	0.12	0.31	**0.45**	1.9	17
CRM; $n = 18$	0.00	0.01	0.05	0.30	**0.63**	3.2	18
JLB; $n = 18$	0.01	0.07	0.11	0.26	**0.55**	3.1	18
CRM; $n = 30$	0.00	0.00	0.03	0.25	**0.72**	5.7	30
JLB; $n = 30$	0.01	0.08	0.11	0.29	**0.51**	5.1	30
SPRT; $n_{tr} = 19$	0.00	0.01	0.02	0.22	**0.75**	5.0	25
Scenario 4, **p**:	0.05	0.05	0.12	0.25	**0.25**		
3+3	0.05	0.12	0.33	0.20	**0.29**	2.2	16
CRM; $n = 18$	0.00	0.02	0.22	0.35	**0.40**	3.6	18
JLB; $n = 18$	0.01	0.11	0.23	0.21	**0.43**	3.2	18
CRM; $n = 30$	0.00	0.01	0.18	0.41	**0.41**	6.2	30
JLB; $n = 30$	0.01	0.10	0.25	0.23	**0.40**	5.6	30
SPRT; $n_{tr} = 19$	0.00	0.02	0.15	0.17	**0.65**	5.9	26

[a]Dose 1 may be concluded unsafe by the 3+3 algorithm, the JLB method, and the stepwise SPRT

4.2 Simulation Results

Table 1.4 show the operating characteristics of the CRM, the JLB method, and the stepwise SPRT based on 5,000 simulation replicates under four dose-toxicity curves. As a benchmark, we also include the results for the 3+3 algorithm that starts at the lowest dose, whereas the others start at dose 3 as in the bortezomib trial. For the CRM and the JLB method, we consider sample size 18 and 30. We did not evaluate the random walk designs due to the memoryless property observed in Sect. 4.1.

The 3+3 algorithm is marked by small sample size, low number of toxicities, and inferior accuracy when compared to the other methods. While it appears safe, the low occurrence of toxicity is due to the fact that majority of the subjects are treated

at low and inefficacious doses. It indicates that the resources invested in phase I trials by the traditional approach are inadequate and misplaced.

The CRM with 18 subjects has good operating characteristics with over 60% probability of selecting the true MTD in Scenarios 1–3, and increasing sample size to 30 improves the accuracy further. Under Scenario 4 where the dose-toxicity curve is flat around the MTD (dose 5), the CRM has mediocre performance, and does not improve as sample size increases. This reveals the fact that the CRM (and almost all other existing methods) assume strict monotonicity of the dose-toxicity curve. This assumption, on the one hand, is true for the classical cytotoxic agents, but on the other hand, needs to be reexamined for therapeutics outside oncology.

The JLB method is uniformly less accurate than the CRM except for Scenario 4, where its performance is mediocre. More importantly, there seems to be a decline in accuracy as sample size increases in Scenarios 3 and 4; it could be because there is a higher likelihood to declare the highest dose as unacceptably toxic with a larger number of "looks" of the data.

The stepwise SPRT has high probability of correct selection in all four scenarios, with average sample sizes about 30. This suggests that an algorithm-based design can be as efficient as the model-based CRM with comparable sample size. There are two caveats, however. First, the sample size of the method is random and can be large; for example, the 90th percentile of the sample size distribution range from 44 to 55 under the simulated scenarios. This can be mended by applying an aggressive truncation (Cheung 2007). Second, the method on average induces slightly higher proportions of toxicity. Safety of the stepwise procedure should also be read in light of the fact that the method recommends any unsafe dose ($p_i \geq 0.45$) for future use with a small probability (<0.10) in comparison to the other methods. Within-trial safety can be further enhanced by choosing a small ε^* in (1.1).

The simulation scenarios here are chosen to illustrate some recurring properties of the designs. In summary, we find that the model-based CRM converges quickly to the right dose even with a small sample size. When more resources are available, the stepwise SPRT is competitive and in fact best when the dose-toxicity curve is flat around the MTD (cf. Scenario 4) because the success of the stepwise procedure does not rely on strict monotonicity of the dose-toxicity curve.

5 Theoretical Properties

General dose-finding theory motivated by ethics and scientific concerns in human trials provides a necessary framework to evaluate the appropriateness of a phase I design. This section reviews two such properties, coherence and consistency, and discusses their relevance.

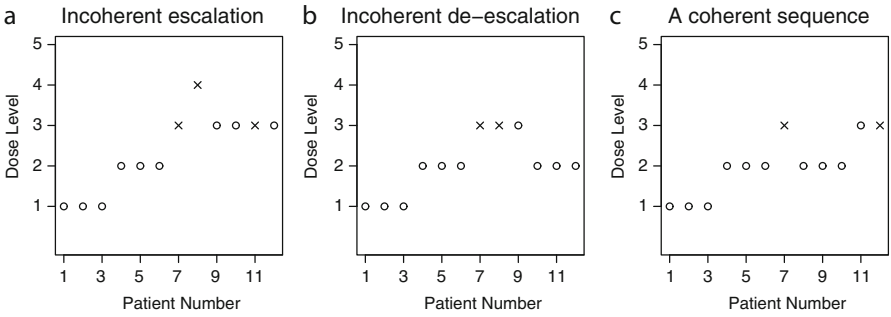

Fig. 1.2 In the outcome sequences, each point represents a patient, with a "o" indicating no toxicity and "x" indicating toxicity. (**a**) Incoherent escalation for patient 7, (**b**) incoherent deescalation for patient 10, (**c**) a coherent outcome sequence

5.1 Coherence

The coherence principles for dose-finding studies state that no escalation should take place for the next enrolled patient if the current patient experiences a toxic outcome, and that de-escalation for the next patient is not appropriate if the current patient shows no sign of toxicity. A dose-escalation design that may induce an incoherent escalation or deescalation is called an incoherent design. Thus, for a coherence design, $\Pr(x_{i+1} > x_i | Y_i = 1) = \Pr(x_{i+1} < x_i | Y_i = 0) = 0$ for all i. From a mathematical viewpoint, coherence is a pointwise property because it involves every possible dose assignment sequence in the sample space. Figure 1.2a shows an outcome sequence with an incoherent escalation, and Fig. 1.2b with an incoherent deescalation; the designs that generate these sequences are incoherent designs. Figure 1.2c shows a coherent outcome sequence; but to establish coherence of the generating design for N patients, all 2^N outcome sequences need to be verified as coherent.

It is almost instinctive to enforce coherence as an ethical criterion in practice. For algorithm-based designs such as the biased coin design where dose decision rules are prespecified, coherence is an intrinsic part of the algorithm. In contrast, coherence of a model-based design needs to be shown on a case-by-case basis. The original CRM (Sect. 3.1.1) has been proved to be coherent in general (Cheung 2005). The two-stage CRM (Sect. 3.1.2) is not necessarily coherent, depending on how the initial design is chosen. Several authors suggest imposing coherence by restriction in a two-stage CRM (Faires 1994; Goodman et al. 1995; Korn et al. 1994). Alternatively, Cheung (2005) provides a coherence condition for the two-stage CRM, and suggests using an initial design sequence that meets the condition. Precisely, let $M_0 = \min\{i : Y_i = 1\}$ denote the transition from the initial design to the CRM in a two-stage design. Then, the coherence condition is met if $\arg\min_k |F(d_k, \hat{\beta}_{M_0}) - \theta| \leq x_{M_0,0}$ where $\hat{\beta}_{M_0}$ is the posterior mean of β given the first M_0 subjects treated according to the initial design. For a trial with N patients, there are only $N-1$ possible transitions. Therefore, we can verify whether

a two-stage CRM is coherent by checking the corresponding outcome sequences for $M_0 = 1, 2, \ldots, N-1$, without the need to go through all 2^N outcomes. The coherence condition may then serve as an objective criterion to choose initial sequence that is compatible with the model-based CRM to produce a smooth transition of stages. Currently, the initial design is typically chosen in an ad hoc manner without clear justification. Since coherence is motivated as an ethical criteria, it may not lead to efficient estimation of the dose-toxicity curve. In the related bioassay literature, efficient designs may sequentially maximize the information (McLeish and Tosh 1990). However, such designs may not be coherent and are not appropriate for human trials.

5.2 Consistency

A model-based design makes use of data available at all doses via a dose-toxicity curve, which if consistently estimated, will lead to a consistent estimator for the true MTD ν. In particular, we are concerned about strong consistency, that is, $x_{n+1} = \nu$ eventually with probability one. At the same time, as the role of the dose-toxicity curve is crucial in a model-based design, it is important to examine the method's robustness against misspecification because there is often not much information to suggest a reliable model at an early clinical stage. The EWOC is consistent under the assumption that $F(x, \beta)$ is a correct model (Zacks et al. 1998), but its robustness under model misspecification has not been examined. The curve-free method suffers no bias due to misspecification, but is inconsistent due to the rigidity discussed in Sect. 3.3. Shen and O'Quigley (1996) prove that the CRM is strongly consistent for ν even when $F(x, \beta)$ is misspecified; Cheung and Chappell (2002) build on the work of O'Quigley and Shen and postulate that the steepness condition (1.3) suffices consistency. Note that condition (1.3) is met if $F(x, \beta)$ is correctly specified, but the converse is not true necessarily. Figure 1.3 shows the probability of correctly selecting dose 5 by models E1 and E2 in Table 1.3 under the scenario $\mu = (0.05, 0.05, 0.10, 0.15, 0.27)$.The CRM with model E1, which satisfies condition (1.3) under μ, performs well at small sample size and improves as sample size increases in accordance with our expectation of consistency. In contrast, the CRM with model E2 does not satisfy condition (1.3) and its performance plateaus as sample size increases. Also, the inconsistent model E2 is much inferior to E1 at small sample size, indicating that asymptotics arguments are predictive of when a CRM model may fail at small sample sizes.

Consistency offers a new perspective about the indifference intervals. Consider the CRM using model E1, whose indifference interval for $\nu = 5$ has a lower limit 0.18. It indicates the model cannot differentiate ν from a lower dose with toxicity rate greater than 18%. Generally, an indifference interval is an interval of toxicity probability to which the CRM will converge. In this sense, indifference interval is an asymptotic concept. While asymptotics may not seem relevant in phase I trials, our experience indicates that models with large indifference intervals (cf. Fig. 1.3)

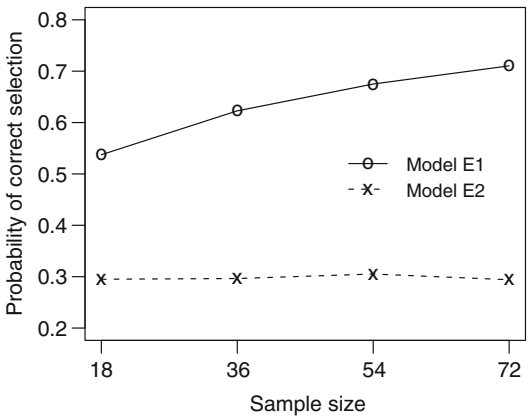

Fig. 1.3 The probability of correct dose selection the CRM using models E1 and E2 in Table 1.3 at various sample sizes. The results were run under the scenario $\mu = (0.05, 0.05, 0.10, 0.15, 0.27)$ based on 5,000 simulation replicates

often does poorly at sample sizes. Finally, the notion of consistency bears an ethical implication that patients enrolled onto the trial will eventually be treated at the right dose, and is thus a practical design feature that encourages patient consent (Rosa et al. 2006).

6 Further Topics

Up to this point, we have assumed that the primary endpoint is a binary toxicity outcome defined over a short-term period. This section discusses how this commonly held assumption may not be practical in the context of the bortezomib trial, and reviews phase I methods tailored for the "nonstandard" study endpoints.

6.1 Delayed Toxicities

In the bortezomib trial, a patient might receive up to six 21-day cycles of treatment, and therefore would be at risk of toxicity throughout the entire treatment period. While it is a common practice to count only toxicity occurring in the first cycle as dose-limiting, we may underestimate the toxicity burden incurred if toxicity tends to occur at a later cycle. On the other hand, it is apparent that awaiting complete follow-up for 126 days before making an escalation decision will cause delays and add administrative burden. In the bortezomib trial and many others, if a toxicity occurs, it may occur at a random time throughout the observation period. As such, it is appropriate to use time-to-toxicity as the study endpoint and define the MTD objective with respect to a prespecified observation window (denoted as T). Cheung and Chappell (2000) propose incorporating the patients' times-to-event into the CRM for dose decisions, and call the method the time-to-event CRM (TITE-CRM).

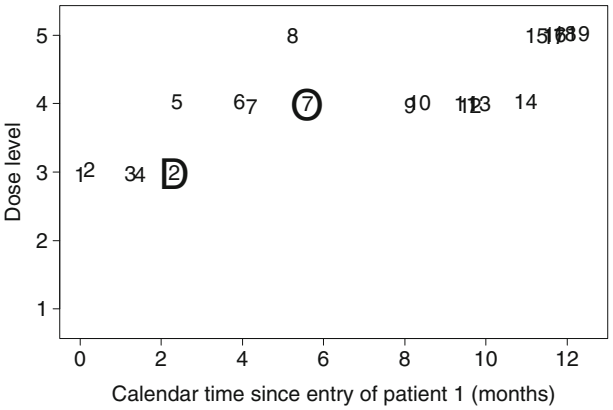

Fig. 1.4 Outcome summary of the bortezomib trial. An unmarked number indicates the patient's entry time. Time of toxicity is marked with a "O," and dropout with a "D." Vertical positions of the numbers are jittered to provide visual separation

The TITE-CRM proceeds in the same manner as the regular CRM: it assumes a one-parameter dose-toxicity model $F(x,\beta)$, starts the trial at the prior MTD, and repeatedly estimates β after every single patient so that the subsequent patient is treated at the updated model-based MTD estimate. A key difference between the TITE-CRM and the CRM is the estimation of β. With the data accrued up to the first n patients, the TITE-CRM estimates β based on a weighted likelihood:

$$\mathscr{L}(\beta)=\prod_{j=1}^{n}\left\{w(C_{j,n+1};T)F(x_j,\beta)\right\}^{Y_{j,n+1}}\left\{1-w(C_{j,n+1};T)F(x_j,\beta)\right\}^{1-Y_{j,n+1}}, \tag{1.4}$$

where $C_{j,n+1}$ and $Y_{j,n+1}$, respectively, denote the follow-up time and toxicity status of patient j just prior to the entry of patient $n+1$. The weight function $w(c;T)$ is increasing in c such that $w(0;T)=0$ and $w(c;T)=1$ for $c\geq T$. Then β may be estimated by $\tilde{\beta}_n=\int\beta\mathscr{L}(\beta)\mathrm{d}G(\beta)/\int\mathscr{L}(\beta)\mathrm{d}G(\beta)$. If a patient is admitted only if the previous patient has been followed completely, then $C_{j,j+1}\equiv 1$, the weighted likelihood (1.4) reduces to the regular binomial likelihood, and $\tilde{\beta}_n=\hat{\beta}_n$.

We may view $w(C_{j,n+1};T)$ as $\Pr(T_j\leq C_{j,n+1}|T_j\leq T,x_j)$ where T_j denote the time-to-toxicity of patient j, so that $\mathscr{L}(\beta)$ is a likelihood based on conditionally independent current status data. In the same spirit of using an underparametrized $F(x,\beta)$, Cheung and Chappell (2000) suggest the use of a predetermined, linear weight function $w(c;T)=c/T$ for $c\leq T$, and demonstrate that the TITE-CRM with such an over-simplistic weight function performs well in many situations.

The TITE-CRM was used for dose-escalation in the bortezomib trial with the linear weight and $T=126$ days. Figure 1.4 displays the trial outcomes. The trial started at dose 3 and the first escalation occurred at day 73 after four patients had shown no toxicity over a fraction of the observation window. The next three patients

were then enrolled at dose 4, and no toxicity was seen in any of the seven subjects by 160 days, at which point the eighth subject arrived and was given dose 5. Shortly after that, patient 7, who had just finished second cycle of treatment, experienced a dose-limiting toxicity. Subsequently, the TITE-CRM assigned dose 4 for the next six patients before reescalated to dose 5, which was the final MTD estimate. This trial enrolled a total of 19 subjects in slightly over a year with only a single dose-limiting toxicity. It indicates the TITE-CRM was reasonably cautious, while avoiding accrual suspensions due to incomplete follow-up. The TITE-CRM have also been applied trials in radiation oncology (Muller et al. 2004; Normolle and Lawrence 2006) and acute stroke (Elkind et al. 2008).

Since the proposal of the TITE-CRM, there have been a number of extensions aiming at improving the method's safety. A general approach is to properly model $w(c;T)$. For example, Cheung (2006) studies the use of adaptive weight function that estimates $w(C_{j,n+1};T)$ with the empirical quantity:

$$\hat{w}_{j,n+1} = u_n \frac{\sum_{i=1}^{n} I\{T_i \le C_{j,n+1}, C_{i,n+1} \ge T\}}{\sum_{i=1}^{n} I\{T_i \le T, C_{i,n+1} \ge T\}} + (1-u_n)\frac{C_{j,n+1}}{T}$$

for some $u_n \to 1$ as n increases. Generally, the use of linear weight appears to be adequate except when the toxicity has a late-onset tendency. In these situations, using adaptive weights prevents erroneous escalation by assigning lesser weights to the early part of the observation window.

A second approach to enhance safety is to temporarily suspend accrual if the accrued data suggest the risk of toxicity at the recommended dose is unacceptably high. Bekele et al. (2008) quantify such risks by formal predictive probabilities. In the bortezomib trial, we would ensure a minimum of 2-week observation period of the last enrolled patient if the next patient was to receive a higher dose. We find this simple approach a practical compromise between trial accrual and patient's safety. For many cancer trials, it is feasible to delay the treatment start date of a patients by 1–2 weeks without administratively closing the trial.

6.2 Toxicity Severity Scores

In the bortezomib trial, toxicity was graded according to the Common Terminology Criteria for Adverse Events v3.0, with "0" indicating no toxicity to "5" indicating toxic death, but the trial objective was defined in terms of a dichotomized outcome. The TITE-CRM was used to find a dose with 25% dose-limiting toxicity rate. This is indeed a common phase I trial practice for which most statistical designs provide a convenient solution. This dichotomy, however, ignores two crucial features of the ordinal toxicity grades. First, grade 4 toxicities such as neuropathy are disabling and irreversible, whereas grade 3 neuropathy are severe and dose-limiting but can be reversed with symptomatic treatment. While 25% grade 3 toxicity rate is tolerable, the tolerance for grade 4 toxicity is much lower. Second, although low-grade toxicity

is not dose-limiting, it signals that the trial has reached near the dosage range where higher grade toxicity becomes likely. The accelerated titration design (Simon et al. 1997) addresses the second issue: it initially escalates after every single patient, and switches to the 3+3 algorithm when one patient has a dose-limiting toxicity, or two patients experience grade 2 toxicity. In addition, the design allow intra-patient escalation if the patient has grade 1 or no toxicity in the previous cycle.

To address the gradation in toxicity severity, the recent statistical literature (Bekele and Thall 2004; Yuan et al. 2007) introduces the concept of severity score for various grades and types of toxicities. Suppose, there are C_j levels of severity for toxicity type j for $j = 1,\ldots,J$, and let w_{jl} denote the elicited severity weight for level l of toxicity type j. Bekele and Thall (2004) define the expected total toxicity burden (TTB) for dose d_k as

$$\psi(\mathbf{w}, d_k, \text{data}) = \sum_{j=1}^{J} \sum_{l=1}^{C_j} w_{jl} E\{\pi_{jl}(d_k, \beta) \,|\, \text{data}\},$$

where $\pi_{ij}(d_k, \beta)$ denotes the probability of experiencing level l of toxicity type j and the expectation is taken with respect to the posterior distribution of the parameter β. The dose-finding algorithm assigns the next dose on the basis of $\psi(\mathbf{w}, d_k, \text{data})$ with respect to a fixed target TTB value ψ^*, and hence requires repeated estimation of $\pi_{jl}(d_k, \beta)$ throughout the trial. The authors use a latent variable modeling approach to estimate $\pi_{jl}(d_k, \beta)$. Because the model is highly parametrized and quite elaborate, it is computationally difficult for model diagnostics and evaluation of the design.

Yuan et al. (2007) subsequently adopt a similar frequentist concept, called equivalent toxicity (ET) score, for the special case with $J = 1$ toxicity type. An ET score for dose d_k is defined as $\psi(d_k, \beta) = \sum_{l=1}^{C} w_l \pi_l(d_k, \beta)$ and the objective is to find the largest d_k with $\psi(d_k, \beta) \leq \psi^*$. Suppose for instance that the respective severity weights for neuropathy at grade 2, 3, and 4 are 0.25, 0.5, and 1.0, and their probabilities of occurrence at an "ideal" dose d_k are 0.1, 0.2, and 0.05. Then $\psi(d_k, \beta) = 0.25 \times 0.1 + 0.5 \times 0.2 + 1.0 \times 0.05 = 0.175$, and we may set $\psi^* = 0.175$. The authors propose to estimate β via quasi-likelihood:

$$\prod_{i=1}^{n} \psi(x_i, \beta)^{W_i} \{1 - \psi(x_i, \beta)\}^{1 - W_i},$$

where $W_i \in \{0.00, 0.25, 0.50, 1.00\}$ is the observed severity weight of patient i, and the model $\psi(x, \beta)$ can be chosen according to the CRM convention (1.2). The quasi-likelihood modeling approach extends directly from the CRM, so calibration of the model ψ can be done in a similar manner to $F(x, \beta)$ as in Sect. 3.1.3.

There are several issues with the use of the severity scores. First, there is no current standard weights assigned to toxicities. Second, in the process of specifying

a target ψ^*, clinicians need to specify an ideal toxicity profile. This is a challenging task as there are virtually infinitely many possible profiles. A third and related point is the fact that very different toxicity profiles may have the same value. For example, a dose that causes grade 2, 3, and 4 neuropathy with probabilities 0.6, 0.03, and 0.01 will have an ET score 0.175. Although this dose has the same ET score as the previous dose, it has low incidence of high grade toxicity and is thus more tolerable than the previous one. Motivated by the first difficulty, Lee et al. (2009) suggest precalibrating the toxicity severity score, which they call toxicity burden score, based on pilot clinical data. To address the other two issues in a sequel, Lee et al. (2001) study an extension of the CRM that defines the trial objective with respect to multiple toxicity constraints applied on the toxicity burden score. Their approach has the advantage that the MTD based on ordinal toxicity is a natural extension of the MTD in the standard settings with a binary toxicity endpoint.

These severity weight approaches provide formal statistical framework that can facilitate the incorporation of ordinal toxicities into dose finding. On the other hand, while they may be implemented at some well-supported research institutions in a specialized and sporadic fashion, methodology and theoretical research directed at these issues is needed before these methods can be broadly implemented.

6.3 Bivariate Designs

Many dose-escalation designs that incorporate both toxicity and efficacy endpoints are motivated by noncancer applications, where safety consideration is relevant but not adequate. Even in cancer trials, the ultimate goal is to identify an efficacious dose for future research. This leads to the combined "phase I/II" clinical trial in which both toxicity and efficacy are considered in the planning stage. The bortezomib trial was one such example with a simple design: the TITE-CRM was used in the phase I portion of the trial where toxicity was the sole basis of dose escalation; an expanded cohort – with proper sample size calculation – was then enrolled to the identified MTD to evaluate the 2-year progression-free survival rate in the study subjects.

For trials with a quick efficacy response, it is feasible to use the response, as well as the toxicity endpoint to choose doses during a trial. Majority of bivariate designs in the statistical literature are model-based and focus on bivariate binary outcomes (Braun 2002; Thall and Cook 2004; Yin et al. 2006). To illustrate their basic idea, let $F_{\mathrm{E}}(x,\beta)$ and $F_{\mathrm{T}}(x,\beta)$ denote the respective probabilities of achieving efficacy and toxicity at dose x, where the parameter β is vector-valued. A dose x is called acceptable if

$$Q_n\{F_{\mathrm{E}}(x,\beta) > \underline{p}_{\mathrm{E}}\} > a_{\mathrm{E}} \text{ and } Q_n\{F_{\mathrm{T}}(x,\beta) < \overline{p}_{\mathrm{T}}\} > a_{\mathrm{T}},$$

where $\underline{p}_{\mathrm{E}}$ and $\overline{p}_{\mathrm{T}}$ are fixed lower and upper limits on the probabilities of efficacy and toxicity specified by clinicians, a_{E} and a_{T} are fixed probability cut-offs, and Q_n denotes probability computed under the posterior distribution given the first n

Table 1.5 Coherence-guided dose decision in bivariate trials

Row	Efficacy	Toxicity	Escalation	Deescalation
1	No	No	Okay	Incoherent: efficacy and toxicity
2	No	Yes	Incoherent: toxicity	Incoherent: efficacy
3	Yes	No	Okay	Incoherent: toxicity
4	Yes	Yes	Incoherent: toxicity	Okay

observations. The basic idea of these approaches is to define a utility $U(x,\beta)$ of dose x as a function of $F_{\mathrm{E}}(x,\beta)$ and $F_{\mathrm{T}}(x,\beta)$; then, among all acceptable doses, the dose with the highest posterior expected utility $E_{Q_n}\{U(x,\beta)\}$ will be selected for use in the next incoming group of patients.

The model-based and utility-based approach need further research in two aspects. First, the impact of underparametrized models on the method's performance needs to be examined. Some (Braun 2002) adopt underparametrized models in light of the robustness of the one-parameter CRM, whereas others (Thall and Cook 2004; Yin et al. 2006) use elaborate models. Although elaborate models have less bias in nonsequential settings, model flexibility may cause undesirable rigidity that confines treatments to suboptimal doses when the doses are assigned sequentially (Cheung 2002).

Second, the preference for the pair $\{F_{\mathrm{E}}(x,\beta),F_{\mathrm{T}}(x,\beta)\}$ is sensitive to how the utility $U(x,\beta)$ is defined. For example, under the efficacy–toxicity odds ratio utility, that is, $U \propto F_{\mathrm{E}}(1-F_{\mathrm{T}})/\{F_{\mathrm{T}}(1-F_{\mathrm{E}})\}$, the pair $(0.8,0.25)$ having an odds ratio 12 is preferred to $(0.2,0.05)$ which has an odds ratio 4.75. The order of preference is reversed under an efficacy–toxicity probability ratio utility, that is, $U \propto F_{\mathrm{E}}/F_{\mathrm{T}}$. Such disagreement in the preference ordering is unavoidable for an utility-based method, because any real-valued function U projects the two-dimensional $(F_{\mathrm{E}},F_{\mathrm{T}})$ onto a one-dimensional space. From a clinical viewpoint, the above two pairs represent very different trade-off preference, and therefore clinician's inputs are crucial in the success of these approaches.

Regardless of the choice of the utility function, the coherence principles for the toxicity-driven designs (Sect. 5.1) can be easily extended to the bivariate setting to simplify dose decisions. Table 1.5 displays all possible outcomes of a patient and the corresponding incoherent moves due to efficacy and toxicity considerations. As defined in the Sect. 5.1, a deescalation is incoherent if the current patient does not experience toxicity (rows 1 and 3 in Table 1.5). Likewise, for efficacy, if the current patient does not achieve a response, it is not ethically sound to treat the next patient at a lower and less efficacious than the current one (rows 1 and 2 in Table 1.5). On the other hand, if the patient has a response, it may be fine to escalate dose for the next patient as long as safety permits it. When the current patient has a toxic outcome but not a response (row 2 in Table 1.5), bivariate coherence implies neither escalation nor deescalation is permissible, and leaves staying at the current dose as the only logical dose decision for the next patient – unless if there is a clear indication that the dose has response rate less than $\underline{p}_{\mathrm{E}}$ and toxicity rate greater than $\overline{p}_{\mathrm{T}}$, then the trial should be closed and the drug concluded futile. These coherence-guided decisions

are transparent and clinically sensible, and provide useful algorithmic restrictions to the complex "black-box" approach of the model-based methods. Coherence may also provide the basic structure for algorithm-based bivariate designs which have received relatively little attention in the literature.

7 Challenge Ahead: Implementation

Much effort is needed to overcome the status quo in phase I trial practice. In this regard, algorithm-based designs are appealing because they are operationally simple and transparent. This being the case, statisticians should still be able to defend a new algorithm-based design and explain why it is superior to the familiar traditional 3+3 method, as it may not be the case at times. In particular, the stepwise procedure has the familiarity advantage because it operates in a similar manner to the 3+3 scheme, while possessing clear statistical properties with prespecified error rates.

On the other hand, our experience with extensive simulations (e.g., Sect. 4.2) indicates that model-based methods such as the CRM provide quick convergence to the right dose, and thus are particularly suited to the "standard" phase I trial setting where the number of subjects are limited. In addition, for the "nonstandard" settings (Sect. 6), model-based approaches have great flexibility to account for variation and the dose-response structure through statistical modeling, and hence provide answers to clinically relevant questions. These advantages, however, are no substitutes for careful planning and operational transparency during the trial.

A crucial step in planning a model-based trial is calibration of the working model, augmented by simulations. As shown in Sect. 5.2, a carefully calibrated CRM model will have robust performance even when the working model is misspecified. For the CRM and its extensions, the R package `dfcrm` (Cheung 2008) provides general model calibration tools and simulation capability to aid trial planning in the R computing environment (R Development Core Team 2008). The R package `dfcrm` also consists of functions to run the CRM, plot an updated dose-toxicity curve, and summarize dose assignments during a trial. For credibility and transparency reasons, it is important for the clinical investigators to see that the method can be implemented in robust and reproducible platform and a comprehensible manner.

Any statistical phase I design should not replace sound clinical judgement. It is almost always appropriate to view the model-based dose assignments as advisory, especially in the event that an escalation is recommended. As there is an increasing trend for constituting a data and safety monitoring committee for early phase cancer trials, the committee may be a natural body to confirm or disapprove the model-based recommendations based on data not formally incorporated in the model. On the other hand, a good model-based design should mimic sound clinical judgment in a *systematic* manner. As a case in point, when a two-stage CRM recommends an escalation when the most current patient has a toxic outcome, we certainly can override the recommendation and impose coherence by restriction. However, the fact that the model-based recommendation does not sound right indicates the model

has not been properly calibrated. As any dubious dose decision by the supposedly superior dose-escalation design shall be effective at convincing the clinicians to stay with the sensible 3+3 algorithm, much care should be given to ensure that clinically sensible judgments can be reproduced in a systematic manner. In this light, there is much room for research and theoretical investigation for the complex model-based designs for the complex "nonstandard" settings, where simulations remain the only tool for method evaluation (cf. Sect. 6.3).

References

Babb J, Rogatko A, Zacks S (1998) Cancer phase I clinical trials: Efficient dose escalation with overdose control. Stat Med 17:1103–1120

Bekele BN, Thall PF (2004) Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. J Am Stat Assoc 99:26–35

Bekele BN, Yuan J, Shen Y, Thall PF (2008) Monitoring late-onset toxicities in phase I trials using predicted risks. Biostatistics 9:442–457

Braun TM (2002) The bivariate continual reassessment method: Extending the CRM to phase I trials of two competing outcomes. Contr Clin Trials 23:240–256

Carbone PP, Krant MJ, Miller SP, Hall TC, Shnider BI, Colsky J, Horton J, Hosley H, Miller JM, Frie E, Schneiderman M (1965) The feasibility of using randomization schemes early in the clinical trials of new chemotherapeutic agents: Hydroxyurea. Clin Pharmacol Therapeut 6:17–24

Cheung YK (2002) On the use of nonparametric curves in phase I trials with low toxicity tolerance. Biometrics 58:237–240

Cheung YK (2005) Coherence principles in dose-finding studies. Biometrika 92:863–873

Cheung YK (2006) Dose-finding with delayed binary outcomes in cancer trials. In: Chevret S (ed) Statistical method for dose-finding experiments. Wiley, NY, pp 225–242

Cheung YK (2007) Sequential implementation of stepwise procedures for identifying the maximum tolerated dose. J Am Stat Assoc 102:1448–1461

Cheung YK (2008) Dose-finding by the continual reassessment method, R package version 0.1-2. http://www.r-project.org

Cheung YK, Chappell R (2000) Sequential designs for phase I clinical trials with late-onset toxicities. Biometrics 56:1177–1182

Cheung YK, Chappell R (2002) A simple technique to evaluate model sensitivity in the continual reassessment method. Biometrics 58:671–674

Cheung YK, Elkind MSV (2010) Stochastic approximation with virtual observations for dose finding on discrete levels. Biometrika 97:109-121

Cheung YK, Gordon PH, Levin B (2006) Selecting promising ALS therapies in clinical trials. Neurology 67:1748–1751

Dixon WJ, Mood AM (1948) A method for obtaining and analyzing sensitivity data. J Am Stat Assoc 60:967–978

Durham SD, Flournoy N (1994) Random walks for quantile estimation. In: Gupta S, Berger J (eds) Statistical decision theory and related topics V. Springer, New York, pp 467–476

Durham SD, Flournoy N, Rosenberger WF (1997) A random walk rule for phase I clinical trials. Biometrics 53:745–760

Elkind MSV, Sacco RL, MacArthur RB, Fink DJ, Peerschke E, Andrews H, Neils G, Stillman J, Corporan T, Leifer D, Cheung K (2008) The Neuroprotection with statin therapy for acute recovery trial (NeuSTART): An adaptive design phase I dose-escalation study of high-dose lovastatin in acute ischeme stroke. Int J Stroke 3:210–218

Faires D (1994) Practical modifications of the continal reassessment method for phase I cancer trials. J Biopharm Stat 4:147–164
Gasparini M, Eisele J (2000) A curve-free method for phase I clinical trials. Biometrics 56:609–615
Goodman SN, Zahurak ML, Piantadosi S (1995) Some practical improvements in the continual reassessment method for phase I studies. Stat Med 14:1149–1161
Ji Y, Li Y, Bekele BN (2007) Dose-finding in phase I clinical trials based on toxicity probability intervals. Clin Trials 4:235–244
Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM (1994) A comparison of two phase I trial designs. Stat Med 13:1799–1806
Lee SM, Hershman D, Martin P, Leonard J, Cheung K (2009) Validation of toxicity burden score for use in phase I clinical trials. J Clin Oncol 27 (Suppl.), 15s (abstr. 2514)
Lee SM, Cheng B, Cheung YK (2001) Continual reassessment method with multiple toxicity constraints. Biostatistics. 12:386–398
Leonard JP, Furman RR, Cheung YKK, Feldman EJ, Cho HJ, Vose JM, Nichols G, Glynn PW, Joyce MA, Ketas J, Ruan J, Carew J, Niesvizky R, LaCasce A, Chadburn A, Cesarman E, Coleman M (2005) Phase I/II trial of bortezomib plus CHOP-Rituximab in diffuse large B cell (DLBCL) and mantle cell lymphoma (MCL): Phase I results. Blood 106:147A–147A
McLeish DL, Tosh D (1990) Sequential designs in bioassay. Biometrics 46:103–116
Muller J, McGinn C, Normolle D et al (2004) A phase I trial using the time-to-event continual reassessment strategy to escalate cisplatin with gemcitabine and radiation therapy for pancreatic cancer. J Clin Oncol 22:238–243
Naylor J, Smith A (1982) Applications of a method for the efficient computation of posterior distributions. Appl Stat 31:214–225
Normolle D, Lawrence T (2006) Designing dose-escalation trials with late-onset toxicities using the time-to-event continual reassessment method. J Clin Oncol 24:4426–4433
O'Quigley J, Shen LZ (1996) Continual reassessment method: A likelihood approach. Biometrics 52:673–684
O'Quigley J, Pepe M, Fisher L (1990) Continual reassessment method: A practical design for phase I clinical trials in cancer. Biometrics 46:33–48
O'Quigley JO, Hughes MD, Fenton T (2001) Dose-finding designs for HIV studies. Biometrics 57:1018–1029
R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.r-project.org
Ratain MJ, Mick R, Schilsky RL, Siegler M (1993) Statistical and ethical issues in the design and conduct of phase I and II clinical trials of new anticancer agents. J Natl Cancer Inst 85:1637–1643
Rosa DD, Harris J, Jayson GC (2006) The best guess approach to phase I trial design. J Clin Oncol 24:206–208
Schneiderman MA (1965) How can we find an optimal dose? Toxicol Appl Pharmacol 7:44–53
Shen LZ, O'Quigley J (1996) Consistency of continual reassessment method under model misspecification. Biometrika 83:395–405
Simon R, Freidlin B, Rubinstein L, Arbuck SG, Collins J, Christian MC (1997) Accelerated titration designs for phase I clinical trials in oncology. J Natl Cancer Inst 89:1138–1147
Storer B (1989) Design and analysis of phase I clinical trials. Biometrics 45:925–937
Storer B, DeMets D (1987) Current phase I/II designs: Are they adequate? J Clin Res Drug Dev 1:121–130
Thall PF, Cook JD (2004) Dose-finding based on efficacy-toxicity trade-offs. Biometrics 60:684–693
Wald A (1945) Sequential tests of statistical hypotheses. Ann Math Stat 16:117–186
Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. Ann Math Stat 19:326–339

Yin G, Li Y, Ji Y (2006) Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratio. Biometrics 62:777–787

Yuan Z, Chappell R, Bailey H (2007) The continual reassessment method for multiple toxicity grades: a Bayesian quasi-likelihood approach. Biometrics 63:173–179

Zacks S, Rogatko A, Babb J (1998) Optimal Bayesian-feasible dose escalation for cancer phase I trials. Stat Probab Lett 38:215–220

Chapter 2
Randomized and Balancing Allocation Schemes for Clinical Trials: Computational Perspectives on Design and Deployment

Vincent J. Carey and Robert Gentleman

1 Conceptual Overview

Randomization of patients to treatments is an essential component of almost all clinical trials. The basic motivation is to allow for appropriate inference to be made and to avoid selection bias that might occur when the treatment likely to be assigned to a given patient is known to the experimenter. Some tension is created by the somewhat contradictory goals of keeping the process from being predictible, while at the same time ensuring that the proportions of patients allocated to the various treatments do not vary considerably from prespecified targets.

A wide variety of allocation schemes have been proposed and many have been used in practice. The monograph of Rosenberger and Lachin (2002) is a very thorough and clear treatment of relevant details on the theory and analysis of randomized designs, and should be consulted by anyone considering doing research on randomization methods. More recently, Rosenberger and Sverdlov (2008) provides updates on research regarding covariate-adaptive randomization methods. Most methods take different approaches to ensuring *balance*, which is the correspondence between the treatment allocations given and those that were intended. In this chapter, we review a number of the basic randomization schemes and provide illustrations using a computer implementation. That implementation is suitable for experimentation and testing the ideas discussed in this chapter.

V.J. Carey
Channing Laboratory, Harvard Medical School, Brigham and Women's Hospital,
181 Longwood Avenue, Boston, MA 02115, USA
e-mail: stvjc@channing.harvard.edu

R. Gentleman (✉)
Bioinformatics and Computational Biology, Genentech, Inc., 1 DNA Way South,
San Francisco, CA 94080, USA
e-mail: rgentlem@gene.com

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7_2,

1.1 Basic Notions for Idealized Experiments

A very simple trial design involves n patients and two treatments, denoted A and B. The trial protocol specifies that N_A patients receive treatment A and $N_B = n - N_A$ patients receive treatment B. There are n allocation events, consisting of recruitment, determination of eligibility and consent, and assignment of treatment to patient i, $i = 1, \ldots, n$. The trial will also involve the determination of a response event or measure, denoted Y_i, for each patient. If $\bar{Y}_t$ denotes the sample mean of the Y_i for those values of i to which treatment t was allocated, then the statistic $\bar{\Delta}_{BA} = \bar{Y}_B - \bar{Y}_A$ is a natural quantity to consider as a measure of the effect of treatment B in comparison to A. Inference can either be based on a known probability distribution for $\bar{\Delta}_{BA}$ under the hypothesis of equal effects if available, or it can be based on permutation-based approaches. Our concern throughout this chapter is to address the question:

- *How should allocations be structured so that estimates of treatment effects and tests of hypotheses about treatment effects will have desirable statistical properties?*

This question includes some inevitably complicated concepts, and our discussion is intentionally limited, avoiding considerations of logistics, ethics, and economics of the experiment. The structure to be imposed on the allocations will involve decision making for each patient, and can also involve consideration of past allocations. Estimates and tests of treatment effects may have different forms depending on scientific objectives of the trial. Statistical properties to be secured may also be application-dependent. We next define a few simple allocation schemes in which we can examine some of these concepts a bit more closely.

Stratification is also dealt with in this section, after all the randomization methods have been described. In our implementation, we found it useful to assign randomizers to strata. This allows for trials where different randomization schemes (in some cases different treatment allocations) can be used within different strata. The approach also greatly facilitates rerandomization since there are often differences in patient populations between strata, suggesting that rerandomization should be carried out separately for each stratum.

1.2 Treatment Allocation Schemes

We now provide brief descriptions of a number of different randomization methods. All methods described here are implemented in the companion software package, *randPack* which is available from the Bioconductor Project. Our discussion focuses on the case where there are two treatments, and often where equal allocation of patients to each treatment is desired. The concepts presented can be extended to more complex situations involving multiple treatments and intentional unequal allocations. In our implementation, the treatment allocation scheme can be represented

as an integer vector, with one number for each of the treatments. For example, a scheme might be represented as `A=2,B=3`, indicating that on average, for every five patients randomized two will be on treatment A and 3 on treatment B. When unequal treatment allocations are desired, we consider imbalance to be the difference between the treatment allocations as given and the ideal (or expected) allocations defined by the randomization scheme.

A major motivation for concern with imbalance is its effect on efficiency of inference on the trial. Under complete randomization, it is possible that substantial imbalance occurs, and only a small number of individuals receive one of the treatments. In this case, uncertainty about the effect of the rarely used treatment can make it very difficult to obtain a clear interpretation of the difference between the treatments. To formalize this, suppose Y_i is the measured response for subject i, $i = 1, \ldots, n$. If the equal-variance two-sample t-test is used to test the null hypothesis of equal mean response values for A and B, the quantity $\bar{\Delta}_{\mathrm{BA}}$ will be divided by its standard error

$$se(\bar{\Delta})[n, n_{\mathrm{A}}, \sigma_{\mathrm{A}}^2, \sigma_{\mathrm{B}}^2] = \sqrt{(1/n_{\mathrm{A}} + 1/n_{\mathrm{B}})(n_{\mathrm{A}} + n_{\mathrm{B}} - 2)^{-1}[(n_{\mathrm{A}} - 1)s_{\mathrm{A}}^2 + (n_{\mathrm{B}} - 1)s_{\mathrm{B}}^2]}$$

and the ratio will be referred to a t distribution with $n - 2$ degrees of freedom. To obtain a simple quantity that illustrates the effect of imbalance (in this case, departure from equal numbers of assignments for A and B) on $se(\bar{\Delta})$, we fix $s_{\mathrm{A}}^2 = s_{\mathrm{B}}^2 = 1$ and obtain $se(\bar{\Delta})[n, n_{\mathrm{A}}, 1, 1] = \sqrt{(n_{\mathrm{A}}^{-1} + n_{\mathrm{B}}^{-1})}$. A dimensionless quantity that will take its minimum value (unity) for a perfectly balanced allocation is the ratio

$$se(\bar{\Delta})[n, n_{\mathrm{A}}, 1, 1] / se(\bar{\Delta})[n, n/2, 1, 1].$$

This ratio is in fact independent of n and can be written as $0.5\sqrt{p^{-1} + (1-p)^{-1}}$, where p is the fraction of allocations to treatment A. It is easy to show that with an imbalance as large as 70:30, the ratio is less than 1.10. Such findings support the general belief that, other things being equal, the risk of efficiency loss owing to imbalances that may arise with pure randomization is not a major concern.

However, overall imbalance is only one of many features controlling the appeal and interpretability of randomized experiments. Because many clinical trial treatments must be administered sequentially, over periods of calendar time that may range into months or years, it becomes important to consider a concept of allocation balance for temporal locales. Even though pure randomization will almost always achieve approximate balance in large enough studies, the problem of "runs" of repeated allocations to a single arm can lead to problems of interpretation. Rosenberger and Lachin (2002) elegantly describe the problem as follows:

> [The occurrence of] severe imbalances at some point during the trial [...] is particularly undesirable if there is a time-heterogeneous covariate that is related to treatment outcome, because imbalances in treatment assignments can then lead to imbalances in those important covariates. [Sect. 3.5]

Simple simulations can be used to learn about features of run lengths in pure randomization. For $n = 50$, the median maximum run length is 6, and the 90th percentile of the distribution of maximum run lengths is 8. So, it would be common in trials of size 50 to encounter a run of assignments to only one arm of size exceeding 10% of the total sample size. The corresponding percentiles for $n = 500$ are 9 and 12, indicating that the fraction of sample size consumed by long runs grows only very slowly with n. Nevertheless, in 25% of trials of size 500, one will see at least two runs of length 10 or more. Such events can cause "data hounds" to salivate profusely. Discovery of the association of such runs with environmental events, for example, can lead to questions of the effects of the imbalance on reliability of interpretation. Other designs, such as the permuted block design, allow strict control over the occurrence of allocation runs.

1.2.1 Complete Randomization

Following Rosenberger and Lachin (2002), we let $T_i = 1$ if the ith patient is given drug A and $T_i = 0$ if the ith patient is given drug B, $i = 1, \ldots, n$. Under complete randomization, the T_i are independent and identically distributed Bernoulli random variables with $Pr(T_i = 1) = 1/2$ for all i. Rosenberger and Lachin also define the quantities $N_A(i) = \sum_{j=1}^{i} T_j, j = 1, \ldots, i$, and $N_B(i) = i - N_A(i)$. Then a signed measure of imbalance for n allocations is $D_n = N_A(n) - N_B(n)$, and a dimensionless measure of imbalance is $Q = N_A(n)/n$. For large n and any $r > 0$,

$$\Pr(|D_n| > r) \approx 2\left\{1 - \Phi\left(\frac{r}{\sqrt{n}}\right)\right\}$$

so that the probability of imbalances exceeding certain thresholds of interest may be computed.

A bit of R code that allows us to explore this concept is as follows. We define a function `percDn(n, p)` which will compute the `p`th percentile of the distribution of $|D_n|$.

```
> tailDn = function(n,r) 2*(1-pnorm(r/sqrt(n)))
> percDn = function(n, p)
+     uniroot( function(r) tailDn(n,r)-p,
+     lower=0, upper=100)$root
> percDn(50, .1)

[1] 11.63087
```

Thus, in about 10% of all completely randomized trials of size 50, there will be an imbalance of at least 11 assignments.

1.2.2 Permuted Blocks

The permuted block design arranges treatment allocations in blocks and the order of assignment is permuted within the block. The number of blocks is then determined by the size of the block and the number of patients to be randomized. In this design balance is guaranteed at the block boundaries within strata (if there are many strata then substantial imbalance can occur due to small amounts in each of the strata). In the permuted block design for a two-arm trial with equal allocation probabilities involving n patients, a block size b is established and $M = n/b$ blocks are constructed. The allocations to treatment are dictated by the sequence of blocks and the assignment sequences within each block, which are obtained by randomly permuting the set of b treatment labels.

There are obvious tradeoffs encountered in adopting a permuted block design in preference to a purely randomized design. First, if the block size is small, it becomes possible to accurately predict future treatment assignments near block boundaries. For this reason, it is often advised to keep the block size as a secret parameter of trial design, or to use randomly varying block sizes. Alternatively, if the block size is large, the problem of runs reemerges. Third, in the "gold standard" multicenter clinical trial design, it is typical to treat center as a stratum. Blocks are then defined within strata, and during the course of the trial, many blocks may be incompletely filled. Interesting analyses of imbalance distributions in stratified permuted block designs will be reviewed in the discussion of stratification below.

Thus, blocked randomization can ensure that allocations to treatment arms are close to balanced with respect to n, the total number of treatments assigned. Avoidance of possibilities of selection bias at block boundaries requires that the block size be kept secret or is randomly varied. When blocking is conducted within a small number of strata, the factors defining the strata will also typically be close to balanced between the arms. If there are many strata, it becomes possible for the trial overall to be substantially unbalanced although most of the strata themselves are individually close to balanced.

1.2.3 Coin- and urn-Based Allocation Schemes

Efron's biased coin design involves a constant parameter $p \in (0.5, 1.0]$, and is denoted as $BCD(p)$. For a two arm study, define D_n to be an increasing function of $N_{\mathrm{A}}(n)$ satisfying $D_N = 0$ when $N_{\mathrm{A}}(n) = n/2$. Allocation for the first treatment occurs using a flip of an unbiased coin; after $j-1$ allocations, the jth allocation will depend on the value of D_{j-1}. Specifically, the jth allocation is to treatment A with probability 1/2 if $D_{j-1} = 0$, with probability p if $D_{j-1} < 0$ and with probability $1-p$ if $D_{j-1} > 0$. The long-run distribution of $|D_n|$ can be obtained with the theory of random walks. Writing $r = p/(1-p)$, the long run probability of perfect balance for n even is $1 - 1/r$; an imbalance of magnitude k occurs with probability decreasing like e^{-k}.

An allocation algorithm that may be regarded as an adaptive biased coin procedure is Wei's urn-based design. For a two-arm study, two parameters specify the procedure: α, the number of balls of types A and B with which the urn is populated at the start of the study, and β, the number of balls of type A (respectively B) added after a draw of a ball of type B (respectively A). The procedure can then be denoted $UD(\alpha, \beta)$. Patients are allocated sequentially according to the type of a single ball that is drawn and replaced; before the next allocation, the β balls of specified type are added. The probability that the first treatment to A is 1/2, and if $N_{\mathrm{B}}(k)$ denotes the number of patients allocated to B among the first k allocations, then the probability that the kth patient is allocated to A is given by $(\alpha + \beta N_{\mathrm{B}}(k-1))[2\alpha + \beta(k-1)]^{-1}$.

Both the biased coin and urn-based allocation procedures engender unconditional probabilities of allocation to treatment A of $1/2$. Rosenberger and Lachin use simulation to show that the variance of the fraction of allocations to A in trials of size 50 is larger for $UD(0,1)$ than for $BCD(2/3)$ (2002, Table 3.4). However, studies of susceptibility to selection bias via optimal guessing of impending allocations shows that the urn design is superior to both $BCD(2/3)$ and permuted block designs; see Chap. 6 of Rosenberger and Lachin (2002).

1.2.4 Minimization

When prognostic factors exist that can affect the tendency of a patient to respond to a study treatment, it is possible in relatively small trials that such factors can become confounded with the treatment assignment. This possibility has led to the creation of a family of allocation procedures known as "minimization." Treatment allocations are made on the basis of the history of past allocations, using the distribution of prognostic factors within arms up to randomization $n-1$. The nth allocation is made to minimize the discrepancy between arms in the distribution of prognostic factors. Rationale for developing such procedures was given by Begg and Iglewicz (1980):

> [B]alancing for prognostic factors prior to the study is necessary as it provides a considerably more efficient comparison of treatments for trials of a typical size .... [W]e believe that trials in which the prognostic factors are seen to be well-balanced will authenticate results in a more convincing manner to a scientific audience than will a sophisticated covariate analysis alone.

Briefly, the idea behind minimization is to attempt to ensure balance on several different variables simultaneously. Once the variables are identified, one must determine how to measure discrepancies from balance, how to combine the imbalance scores into a single score for each potential treatment allocation and finally there must be an algorithm to select the treatment given these combined scores.

To illustrate minimization in concrete terms, consider the following example. Patients are allocated to one of two treatments, A or B. Two factors, sex and stage (at three levels) were chosen as factors for which balance is desired. A new male patient with stage I disease needs to be entered and the current state of treatment allocations is given in Table 2.1.

Table 2.1 Treatment allocations for minimization example

		A	B
M	I	3	2
	II	1	4
	III	3	1
F	I	2	2
	II	3	3
	III	1	1

For sex, the imbalance will be measured as the absolute value of the difference between arms in count of males, and for stage it will be the square of the difference between arms. The selection of the imbalance measure (absolute value, square of the difference, etc.) is a user defined choice. We first determine the state of the experiment for males. There are 14 males enrolled and of these seven are on treatment A and seven on treatment B. Now, for each treatment, we assign the new patient to it and then measure the mismatch score. So, first with A, we get $\text{abs}(8-7)=1$ and then with B, we get $\text{abs}(7-8)=1$. So there is no preference.

Next, we consider those with stage I. There are nine enrolled patients with stage I of the disease; 5 on A and 4 on B. If the new patient is assigned to A, we get a score of $(6-4)^2=4$ and if assigned to B, we get $(5-5)^2=0$.

Now, we need some means of combining the scores on the two factors. A simple way to do this is to use weights. So, lets say that stage is really important and we will give it weight 4 while sex is less important and will be given weight 1. Thus, the overall score if the new patient is assigned to A is 17 while if the patient is assigned to B the score is 1. Now, we need to decide which assignment to make. There are many possibilities. Examples include: (1) choose the treatment with the lowest score; (2) choose the one with the lowest score with some fixed probability; (3) choose a treatment with probability proportional to score.

To briefly summarize, one must determine which factors to use. Then, for each of these factors one must determine what imbalance measure to use. Then, one must determine how to combine those scores, a common choice is a weighted average. And finally, once scores are computed for each potential treatment allocation then an algorithm to select one must be specified. In our implementation, we used a deterministic method to select the treatment; the treatment with the best score is chosen with probability 1. An unfortunate effect of this choice is that for a fixed set of patients once the first patient has been assigned the remainder of the assignments are essentially deterministic. Making use of some form of randomization at this stage is likely to be important in practice since it makes the process less deterministic.

The method has several potential advantages in that it is able to provide balance across many covariates. However, it has the substantial disadvantage of being somewhat difficult to assess the performance using classical statistical methods. Hence, for this allocation scheme in particular access to inference via rerandomization or simulation seems particularly important.

1.3 Stratification

In a very thorough review of randomization in clinical trials +design, Zelen (1974) wrote:

> One of the key dictums in experimentation is to take account of all known factors which may significantly affect the outcome of a trial. Not to do so may introduce biases in the data, which may lead to drawing wrong conclusions and possibly introduce so much variability in the data so as to completely obscure any real differences among the treatments. When the factors influencing response are known, we can take this into account in the initial randomization. Then the randomization is referred to as stratified randomization. For example, some factors important for planning cancer studies are: institution, anatomical staging, histological type, prior treatment, general health of patient, demographic factors, etc.
>
> Taking account of these other factors in the initial treatment assignment ensures that each of the therapies has an equal distribution of patients with regard to the important characteristics which may significantly affect response. Of course, using a stratified randomization scheme increases the bookkeeping of the clinical trial. One must weigh the gain in efficiency of the trial against the increased complexity of running the study.

In line with these concluding caveats, a challenge to the desirability of stratified designs was raised by a group of highly eminent statisticians writing in 1976:

> The improvement in the sensitivity of a clinical trial to be expected from achieving perfect balance between the numbers on each treatment in each retrospective stratum, instead of letting them be defined by chance, is just that to be expected from randomizing a single extra patient into each retrospective stratum.... However many initial strata are defined, only a few retrospective strata will be needed, and so the expected benefits from initial stratification are slighter than would intuitively be expected; indeed, if the organizational complexity of stratification at the time of randomization reduced collaboration at all, a net loss of efficiency would be the likely result Peto et al. (1976).

These authors conclude with the caveat that multicenter trials should be stratified by center.

In practice, it is seldom the case that investigators have only one or two factors of concern for which stratification is proposed, and discussions of the cost-effectiveness of the construction of numerous strata are common in trial design considerations. Chapter 8 of Rosenberger and Lachin (2002) reviews classical developments in reasoning about the efficiency of stratified randomization in the context of stratified analyis. It is shown that in large trials, efficiencies of post-hoc stratified analyses are not substantially increased for trials in which stratified (as opposed to simple) randomization was used for treatment allocation. In small trials, stratified randomization does yield efficiency benefits, but the number of strata that can be feasibly used in small trials is often quite small.

The implementation of stratified randomization is not conceptually complex. Each stratum has its own allocation algorithm. Logistical concerns arise when determination of stratum membership is uncertain (as it may often be when “race” is used as a stratum), or when members of certain strata are hard to recruit. Randomized trials with stratified randomization can be unbalanced when numerous strata are all slightly unbalanced in the same way. Approximate distributions for imbalance measures under models for the occurrence of unfilled blocks are derived by Hallstrom and Davis (1988).

1.4 Inference Procedures: Rerandomization

Data from clinical trials are generally analyzed using a population model-based approach. But there are some concerns with taking such an approach, primary among them the fact that randomization was restricted and that certain patterns of treatment allocation were not possible. As noted by both Efron (1971) and Cox (1982), the likely effect of ensuring close balance at all times is a reduction in variance, making population-based analyses (or ones based on complete randomization) conservative. While some effects may be alleviated in relatively large samples, it is not uncommon to perform subset analyses (say on particular centers) and these analyses would likely benefit from an approach that was more faithful to the randomization mechanism employed in the study. We consider such *rerandomization* tests below.

Permutation tests are a widely used method for testing hypotheses. The approach is quite simple: under the null hypothesis that the treatment has no effect then the distribution of the trial test statistic is identical for all assignments of treatment labels to patients. The null distribution of the trial test statistic can be computed by permuting the treatment labels against the observed outcomes. For each permutation the test statistic of interest can be computed and the observed value of that statistic, for the labels as originally assigned, is compared to the permutation-based reference distribution. We note that in general there is no real need to enumerate all possible permutations and that rather, a Monte Carlo approach, where a large number of possible permutations is sampled, should provide sufficient accuracy.

One might argue that not all permutations of the treatments are valid, given the randomization scheme, and while that is indeed true, one can just as easily permute the arrival times of the patients, and treat the treatments as fixed. Clearly the same distribution achieves, and hence provided that one can rationalize permuting the arrival times of patients (or the order in which they are assigned treatments) then permutation-based inference has validity. However, in many trials, especially those that have long accrual times, such as many cancer trials, there is temporal heterogeneity in the patient population. Thus, one may not be able to obtain valid inference in this way.

Rerandomization is the process of keeping the patients ordered as they were observed in the trial and generating new treatment allocations according to the randomization scheme that was chosen for the trial. For example, if permuted blocks were used, then one could generate many different sets of permuted blocks, and apply each set in turn, to develop the necessary reference set.

It is worth pointing out an important difference between permutation-based inference and rerandomization. With permutation-based inference, we are conditioning on the marginal distribution of the treatment assignments. From this a reference set of test statistics, consistent with that conditioning are generated. For rerandomization, we reassign treatment allocations according to the protocol and can get any set of allocations consistent with the protocol. The rerandomization

approach, thus, seems to be more aligned with the question of what would happen if this same set of patients was reassigned (under the null hypothesis of no treatment affect).

1.5 Exceptions to the Ideal Setting

In practice, clinical trials can be very messy affairs. Treatments may become unavailable for some period of time; patient information may be improperly entered and manual intervention is needed; sites enter and leave the study, perhaps due to lapses in IRB approval. A comprehensive system would need to include methods for verifying patient, and institution, eligibility. It would need to allow treatments to be suspended and resumed.

When there are many strata, it is possible for the trial to be quite unbalanced when all strata are considered, but for no strata to be very unbalanced. In some implementations additional constraints might be added, for example, to ensure that the allocations at all institutions or sites are approximately balanced. We will not directly consider such mechanisms since they typically add a great deal of complexity to the system.

Another important factor in the implementation of any system is predictibility. If there is a chance that the individual randomizing patients can readily guess what the next treatment will be, then there is some chance that bias will creep in. They may be less willing to enroll a patient in the trial if they think that the patient is poorly suited to a treatment that is highly likely to be the next treatment allocated.

2 Computational Architectures for Allocation in Clinical Trials

In this section, we describe and illustrate features of systems that can perform treatment assignments for clinical trials in reasonably flexible ways. Some important properties of such systems include

- *Standardized formal specification.* The description of a trial is typically provided in the form of a protocol document composed by and for human readership. These are invaluable and irreplaceable but are almost always ambiguous in important ways. Many aspects of trial design and implementation are algorithmic in nature and can be specified through computer programs whose soundness and validity can be mechanically checked. A more effective approach to trial description will consist of a combination of textual protocol and computable specification components.
- *Validity criteria for data elements.* It is often possible to prospectively define aspects of trial data that must be satisfied if the protocol is correctly implemented. Examples important for randomization include form and content of enrollment

variables, particularly those involved in eligibility and stratification determinations. Allocation tools should have access to and make use of validity conditions as a matter of course.

- *Modular design.* For any given clinical trial network, a universal backbone of textual and software components should be complemented by components that can be interchanged to vary trial operations as desired. For example, the representation of a stratum should allow specification of an allocation procedure specific to that stratum. Conditions for determining patient eligibility and making endpoint decisions should be defined in modules that can be "plugged in" to the trial backbone to flexibly define the trial's operations. When modularity principles are followed, trial simulation can be used to examine impacts of using different procedures in different stages of trial conduct.
- *Auditability.* Each allocation decision needs to be reconstructible on demand so that assurances of unbiasedness and soundness can be made in monitoring or trial review.

There are many ways of designing computing systems that implement clinical trials allocation procedures. We will use the R programming language to define data structures and functional workflow components that create trial data and compute treatment allocations. We do not want to get into the details of the underlying software implementation but provide here a concise conceptual framework for our implementation. Various entities, such as a randomizer or a clinical trial can be represented as objects, with specific slots for different entities. The process of initiating a clinical trial and assigning patients requires the creation of an object that now contains data, in addition to the specifications of the trial. To achieve this we define one class, `ClinicalExperiment` that holds the specifications and a second class, `ClinicalTrial` that holds the experiment specifications, the patient data and the randomizers.

We have taken a similar approach in our design of the randomizers. There are two basic classes, `RandomizerDesc`, which provides the description of a randomizer and `Randomizer` which holds an instance of the randomizer for a particular trial or strata within a trial. Currently, we support five types of randomization, random, permuted blocks, urn models, minimization, and Efron's biased coin.

2.1 Experiment- and Patient-Level Data Structures

Here, we formalize the construction of a clinical experiment object and patient data that can be used to populate it. We want to have sufficient structure to be interesting and realistic. We create a trial randomized at four centers (`C1, C2, C3, C4`), with three covariates, sex, age (grouped), and a random standard Normal variate (the outcome, hence, our trial is null since the same distribution of outcomes is used for all individuals). The allocation will be stratified by center, age and sex will be used for covariate allocations.

First we create a set of patient identifiers, these are different for each of the centers. As a technical aside, the use of the symbol `L` after integers, for example, `1000L`, indicates to R that they are to be treated as integer constants.

```
> library(randPack)
>   #  establish form and limits of IDs
> pidStruct = new("PatientID",
+         strata=c("C1", "C2", "C3", "C4"),
+         start = c(1000L, 2000L, 3000L, 4000L),
+         stop = c(1999L, 2999L, 3999L, 4999L))
> pidStruct

randPack PatientID instance for  4  strata.
The total number of IDs prepared is 4000.
```

Now, we specify factors relevant to randomization and the treatment names and balanced allocation ratio (we used `A=2, B=2` to get a block size of four when permuted blocks is implemented):

```
> ce = new("ClinicalExperiment",
+            name="bookdemo",
+            factors=list(
+              center=c("C1", "C2", "C3", "C4"),
+              sex=c("Male", "Female"),
+              age = c("a50-54", "a55-59", "a65-69"),
+              stdnor = "numeric"),
+            treatments=c(A=2L, B=2L),
+            randomization=list(),
+   # following operates on PatientData instance
+            strataFun = function(x)x@strata,
+            patientIDs = pidStruct)
> ce

ClinicalExperiment:  bookdemo
  With 2 treatments
         A B
  With 0 strata
```

In principle, information in the `factors` component specified above could be used to formalize eligibility conditions for patients considered for enrollment. Note that since we have not yet specified our strata, the value is currently set to `0`. Once we have instantiated the trial (through a call to `createTrial` below), the number of strata will be set, and hence printed.

2.2 Randomization Specification

We use *randPack* class definitions to construct a permuted blocks randomizer. Note that we do not need to specify the block size as that is inferred from our description of the experiment above.

```
> pbdesc = new("PermutedBlockDesc",
+         treatments = ce@treatments,
+         type="PermutedBlock",
+         numBlocks=250L)
```

The following code associates a permuted blocks method with each center:

```
> randomization(ce) = list(C1=list(pbdesc),
+                 C2=list(pbdesc),
+                 C3=list(pbdesc),
+                                 C4=list(pbdesc))
```

We instantiate the associated clinical trial using the `createTrial` function. We specify a seed for the random number generator separately for each strata using the `seed` argument.

```
> CT1 = createTrial(ce, seed = c(301, 401, 501, 601))
> CT1

randPack Clinical Trial instance
ClinicalExperiment:  bookdemo
  With 2 treatments
         A B
  With 4 strata
         C1 C2 C3 C4
  Randomizer:  C1
    Patients Randomized 0
  Randomizer:  C2
    Patients Randomized 0
  Randomizer:  C3
    Patients Randomized 0
  Randomizer:  C4
    Patients Randomized 0
```

2.3 Cohort Simulation: Allocations

Since we wrote the function `simPats`, we can simulate a cohort of enrolees with a simple list-based schema. We specify the proportions of enrolees from different centers, proportions of factors expected, and distributions of continuous

covariates, each of these values is generated independently of the others. Statistical dependencies among the various factors schematized here could be incorporated in more realistic simulation frameworks by modifying the `simPats` function.

```
> coh1 =  list(center=c(C1=.4, C2=.2, C3=.1, C4=.3),
+              sex=c(Male=.5, Female=.5),
+              age = c('a50-54'=.4, 'a55-59'=.3,
+                      'a60-64'=.1, 'a65-69'=.2),
+              stdnor = function(x)rnorm(x))
> simDat = simPats(100, coh1)
> simDat[1:5,]

  center    sex    age     stdnor
1     C1   Male a65-69  2.2232335
2     C4   Male a50-54 -1.0025901
3     C2 Female a65-69  0.1826401
4     C1   Male a60-64  0.6004586
5     C1 Female a65-69  0.6196092
```

We obtain the allocation for patients from this cohort as follows. First, we create 100 `PatientData` containers on the basis of the cohort table generated above.

```
> pdlist = lapply(1:100, function(i)
+  new("PatientData", date=Sys.Date(), name=
+     paste("Patient", i, sep=""),
+     covariates=simDat[i, c("age", "sex", "stdnor")],
+     strata=as.character(simDat[i, "center"])))
```

Now we invoke the `getTreatment` method for the trial on each of the patients:

```
> trts = lapply(1:100, function(i) getTreatment(CT1,
+                         pdlist[[i]]))
> CT1

randPack Clinical Trial instance
ClinicalExperiment:  bookdemo
  With 2 treatments
         A B
  With 4 strata
         C1 C2 C3 C4
  Randomizer:  C1
    Patients Randomized 43
  Randomizer:  C2
    Patients Randomized 24
  Randomizer:  C3
    Patients Randomized 5
  Randomizer:  C4
    Patients Randomized 28
```

The covariates along with the allocations are obtained with the code

```
> ei = getEnrolleeInfo(CT1)
> names(ei)

[1] "C1" "C2" "C3" "C4"

> ei[["C1"]][1:4, ]

       name    age    sex    stdnor alloc
1  Patient1 a65-69   Male 2.2232335     B
2  Patient4 a60-64   Male 0.6004586     B
3  Patient5 a65-69 Female 0.6196092     A
4 Patient14 a50-54   Male 1.3650090     A
```

generating one data frame per stratum. In most of the analyses reported below, we collapse this to a single dataframe and do not report stratum specific values, although in practice one would perform stratum specific analyses.

To conclude this sketch of allocation computations, we illustrate how a deterministic minimization procedure can be used with the same patient data. The *randPack* package includes functions that implement the allocation criteria of Taves (1974) and Pocock and Simon (1975). We will use the latter with the objective of minimizing age imbalance between arms.

We begin by defining a minimization randomizer and then instantiate a new clinical experiment object that uses this randomizer. Random number generator seeds are supplied to deal with the cases in which a randomized choice is required.

```
> md = new("MinimizationDesc", treatments=c(A=1L,
            B=1L),
+    method=minimizePocSim, type="Minimization",
+    featuresInUse="age")
> dd = list(C1=list(md), C2=list(md),
+               C3=list(md), C4=list(md))
> randomization(ce) = dd
> CT2 = createTrial(ce, seed=c(301,401,501,601))
```

Allocations are made in the same way as above, by calling the `getTreatment` function.

```
> trtsMin = lapply(1:100, function(i)
+        getTreatment(CT2, pdlist[[i]]))
> CT2

randPack Clinical Trial instance
ClinicalExperiment:  bookdemo
  With 2 treatments
         A B
  With 4 strata
```

```
        C1 C2 C3 C4
Randomizer:  C1
  Patients Randomized 43
Randomizer:  C2
  Patients Randomized 24
Randomizer:  C3
  Patients Randomized 5
Randomizer:  C4
  Patients Randomized 28
```

A comparison of the association of age and allocation in trials CT1 and CT2 can be conducted using χ^2 statistics. First, we obtain the statistic for the permuted block assignments.

```
> ct1dat = do.call(rbind, getEnrolleeInfo(CT1))
> with(ct1dat, chisq.test(table(age, alloc)))

Pearson's Chi-squared test

data:  table(age, alloc)
X-squared = 1.1077, df = 3, p-value = 0.7752
```

And then repeat the computation for the minimization assignments.

```
> ct2dat = do.call(rbind, getEnrolleeInfo(CT2))
> with(ct2dat, chisq.test(table(age, alloc)))

Pearson's Chi-squared test

data:  table(age, alloc)
X-squared = 0.0361, df = 3, p-value = 0.9982
```

While neither trial is threatened by a concern that treatment and age are confounded, the allocations via minimization yield marginal frequencies that are more consistent with statistical independence of treatment and age than those obtained via permuted blocks.

2.4 Summary on Computational Infrastructure

In the preceding subsections, we have shown a few of the key components required to support flexible treatment allocations in randomized clinical trials. The structure of the experiment is defined by the target treatment allocation ratios and the stratification factors. Trial allocation procedures are specified on a per-stratum basis and can employ arbitrary information on patient characteristics to compute allocations.

Table 2.2 Summary of functions in *randPack*

createTrial	Create an instance of the ClinicalTrials class
factorNames	Access the factor names in a ClincialExperiment instance
getEnrolleeInfo	Returns one dataframe per strata of enrollee information
getTreatment	Performs the assignment of a treatment to the supplied patient
makeRandomizer	Instantiate the randomizer for a strata
minimizePocSim	The Pocock–Simon minimizer
minimizeTaves	The Taves minimizer
numberOfFactorLevels	Returns a list of the factors and how many levels each has
numberOfTreatments	Returns the number of treatments specified for a clinical trial
`randomization<-`	Assignment of the randomizer to a strata in a clinical trial
simPats	Simple function to simulate patient populations
treatmentFactors	Lists all factors associated with the experiment
treatmentNames	Lists the names of the treatments associated with the experiment

The *randPack* functions used for illustration in this section would require strengthening for use in real clinical applications. We have not addressed the tasks of validating patient eligibility or verifying the logical and substantive soundness of patient covariate information. The problem of communicating allocations to remote sites upon request requires attention to possible conflicts when requests are made simultaneously; a genuine relational database system with record locking should be used as the "back end" to any serious system for clinical trial allocation infrastructure. Finally, attention to concerns about information security and effectiveness of masking of treatment assignments in blinded trial designs is important but beyond the scope of this chapter.

Despite these limitations, the implementation in *randPack* provides readily accessible tools for exploration and assessment of different approaches to treatment allocation, and we illustrate some facets of this facility in the next section.

We summarize the set of functions provided by *randPack* in Table 2.2.

3 Evaluating Allocation Methods

We begin a brief comparative investigation of allocation methods by constructing a variety of randomizer description objects and creating a tool to generate allocations according to all randomizer types for a fixed patient dataset. We will reuse the ClinicalExperiment object defined above in Sect. 2.1 for illustration.

Then, we create a function with parameters `NSUB` (population size) and `popschem` (population structure) that generates a single realization of a cohort from the schematized population, and defines five trials based on this cohort. Each trial uses a different allocation algorithm.

```
> getTrialSet = function(NSUB=100, popschem) {
+   simDat = simPats(NSUB, popschem)
```

```
+
+   pdlist = lapply(1:NSUB, function(i)
+     new("PatientData",
+          date=Sys.Date(), name=paste("Patient", i,
    sep=""),
+     covariates=simDat[i, c("age", "sex", "stdnor")],
+     strata=as.character(simDat[i, "center"])))
+
+   ai = as.integer
+   rdesc = new("RandomDesc", treatments
     = ce@treatments,
+        type = "Random", numPatients = ai(NSUB))
+
+   pbdesc = new("PermutedBlockDesc",
+                  treatments = ce@treatments,
+                  type="PermutedBlock",
+                  numBlocks=ai(NSUB/
    sum(ce@treatments)))
+
+   ebcdesc = new("EfronBiasedCoinDesc",
+        treatments = ce@treatments,
+        type="EfronBiasedCoin", numPatients=ai(NSUB),
    p=2/3)
+
+   urndesc = new("UrnDesc", treatments
     = ce@treatments,
+        type="Urn", numPatients=ai(NSUB), alpha=0,
    beta=1)
+
+   mdesc = new("MinimizationDesc", treatments
    =ce@treatments,
+    method=minimizePocSim, type="Minimization",
+         featuresInUse="age")
+
+   dlist = list(rand=rdesc, permBl=pbdesc,
    biasedCoin_66=ebcdesc,
+     urn_0_1=urndesc,
+     pocSimMin=mdesc)
+
+  d2strata = function(x) list(C1=list(x), C2=list(x),
+          C3=list(x), C4=list(x))
+
+   allDescs = lapply(dlist, d2strata)
+
+   allTrials = lapply(allDescs, function(x) {
```

```
+       randomization(ce) = x
+       createTrial(ce, seed=c(301,401,501,601))
+       })
+    list(allTrials=allTrials, pdlist=pdlist)
+ }
```

We have now described five different trials. They are all using the same patient data, so we have a form of blocking for our simulation experiment. We next run the `getTrialSet` function to instantiate the designs for the trials.

```
> oneSetup = getTrialSet(NSUB = 100, popschem = coh1)
```

The allocations prescribed by each design are assigned in the statement below (the call to `getTreatment` causes the actual assignment).

```
> allEnrollments = lapply(oneSetup$allTrials,
    function(x) lapply(1:100,
+     function(i) getTreatment(x,
    oneSetup$pdlist[[i]])))
```

Now, we can extract the assignments and carry out any analysis of interest on them.

```
> allDF1 = lapply(oneSetup$allTrials, function(x)
    do.call(rbind,
+     getEnrolleeInfo(x)))
```

We can tabulate the ultimate allocation counts for each method easily:

```
> sapply(allDF1, function(x) table(x$alloc))
```

```
    rand permBl biasedCoin_66 urn_0_1 pocSimMin
A   47     50            51      52        48
B   53     50            49      48        52
```

3.1 Departures from Target Allocation Ratios

We need a large number of realizations of the previously described setup to check for systematic differences among trial designs in achievement of target allocation ratios. The following code generates 100 realizations.

```
> if (!exists("allDF")) {
+  if (file.exists("allDF.rda")) load("allDF.rda")
+  else allDF = lapply(1:100, function(i) {
+  cat(i)
+  setup = getTrialSet(NSUB=100, popschem=coh1)
```

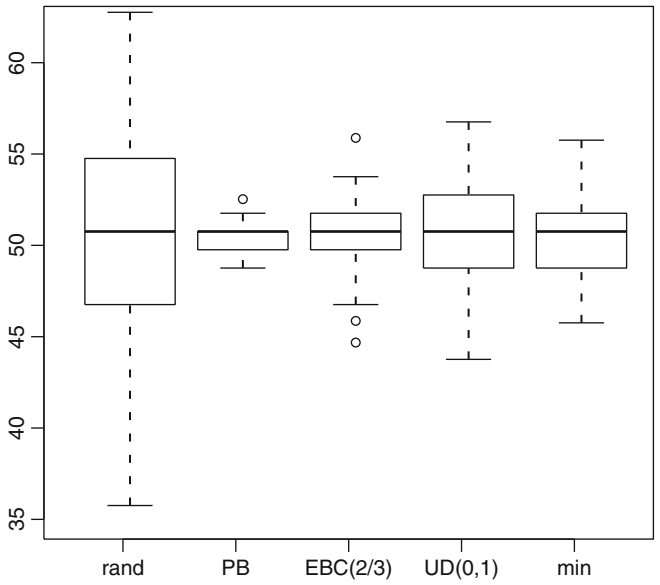

Fig. 2.1 Distributions of numbers of assignments to treatment A in 100 nominally balanced trial realizations for 100 patients in 4 strata as defined in object `ce` (Sect. 2.1). Notation: rand denotes purely random allocation; PB denotes permuted blocks, fixed blocksize of 4; EBC(2/3) denoted Efron's biased coin with $p = 2/3$; UD(0,1) is Wei's urn design with $\alpha = 0$, $\beta = 1$, and min is Pocock–Simon minimization

```
+  tmp = lapply(setup$allTrials, function(x)
+                 lapply(1:100, function(j)
+                 getTreatment(x, setup$pdlist[[j]])))
+  lapply(setup$allTrials, function(x) do.call(rbind,
+        getEnrolleeInfo(x)))
+  })
+  save(allDF, file="allDF.rda")
+ }
> tabulateA = sapply(1:100, function(i)
+    sapply(1:5, function(x)sum(allDF[[i]][[x]]$
      alloc=="A")))
> ac = as.character
> tabulateRuns = sapply(1:100, function(i)
+    sapply(1:5, function(x)
+      max(rle(ac(allDF[[i]][[x]]$alloc))$length)))
```

Figure 2.1 shows the distribution of the number of allocations to treatment A over 100 realizations for each design. Pure randomization shows the greatest degree of fluctuation; permuted blocks fails to achieve perfect allocation owing to incomplete blocks in strata.

We now consider the distribution of lengths of longest runs of allocations. We simply concatenate the allocations across strata within each realization. Figure 2.2 shows that permuted blocks allocations confine maximal run lengths very effectively, while some long runs are observed in all other procedures.

```
> ac = as.character
> tabulateRuns = sapply(1:100, function(i)
+    sapply(1:5, function(x)
+        max(rle(ac(allDF[[i]][[x]]$alloc))$length)))
> numrun = as.numeric(t(tabulateRuns))
> type = rep(c("rand", "PB", "EBC(2/3)",
+        "UD(0,1)", "min"), each=100)
> ftype = factor(type)
```

3.2 Avoiding Accidental Confounding

The minimization procedure is designed to ensure that trial arms do not exhibit substantial differences in the distribution of specific cofactors. We have implemented minimization to work with categorical cofactors, and in the simulation study we have used a discrete representation of patient age for the minimization algorithm. Figure 2.3 shows the effectiveness of the minimization algorithm in forcing the distributions of treatment and age to be approximately statistically independent, in that the χ^2 statistic measuring departure of observed frequencies from those expected under independence is generally much smaller for minimization allocations than for any other procedure.

3.3 Inference: Permutation vs. Rerandomization

In this subsection, we consider relationships among simulation-based distributions of test statistics obtained with different allocation algorithms. The response variable is a standard Gaussian deviate generated independently for each individual at time of cohort construction. The statistic of interest is the standard error (sampling standard deviation) of the estimated mean treatment effect. Three classes of statistic are of interest:

- *Model-based*: The standard error of the t-statistic for the two-sample test of a treatment effect on the mean response.
- *Permutation-based*: The empirical standard deviation of the permutation distribution of the estimated mean difference.

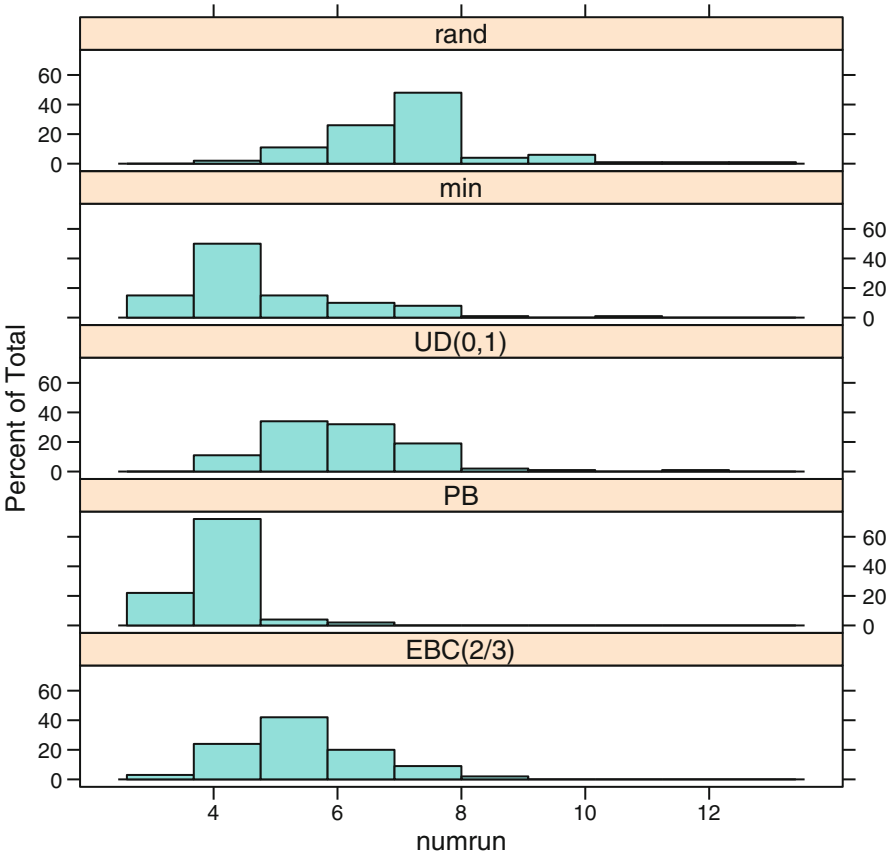

Fig. 2.2 Distributions of maximum run length over 100 trial realizations under various randomization schemes. Notation: rand denotes purely random allocation; PB denotes permuted blocks, fixed blocksize of 4; EBC(2/3) denoted Efron's biased coin with $p = 2/3$; UD(0,1) is Wei's urn design with $\alpha = 0$, $\beta = 1$, and min is Pocock–Simon minimization

- *Rerandomization-based*: The empirical standard deviation of the rerandomization distribution of the estimated mean difference.

  The model-based statistics are generated as follows:

```
> alls = sapply(allDF, sapply,
+     function(x)
+        summary(lm(x$stdnor~x$alloc))$coef[2,"Std.
    Error"])
```

  For permutation-based inference, we have an inner iteration.

```
> NPERM = 100
> i = 0  # for monitoring
```

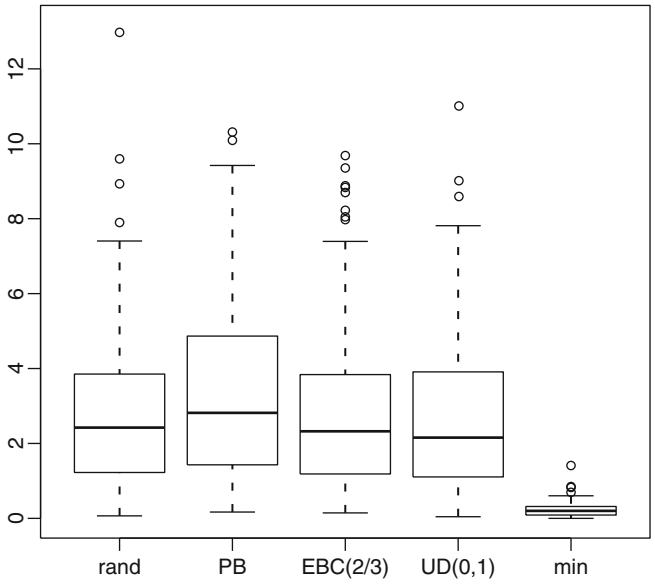

Fig. 2.3 Distributions of χ^2 statistics testing for independence between enrollee age and treatment allocation. Notation: rand denotes purely random allocation; PB denotes permuted blocks, fixed blocksize of 4; EBC(2/3) denoted Efron's biased coin with $p = 2/3$; UD(0,1) is Wei's urn design with $\alpha = 0$, $\beta = 1$, and min is Pocock–Simon minimization

```
> perm = function(x) {i<<-i+1; cat(i);
+      sample(x, size=length(x), replace=FALSE)}
> permstat1 = function(w)
+      summary(lm(w$stdnor~perm(w$alloc)))$coef[2,
         "Estimate"]
> permsd = function(w)
+      sd(sapply(1:NPERM, function(z) permstat1(w)))
> allsPerm = sapply(allDF, sapply,
    function(x)permsd(x))
```

Figure 2.4 shows that for this application, model-based and permutation-based inferences will be essentially indistinguishable. This is to be expected as the test used is optimal for the data generation scheme.

Computation of rerandomization distributions of test statistics of interest is relatively simple given the modularity of the design of the system. The following code chunk defines a function that takes a list of stratum-specific randomizer descriptions and a clinical experiment object as input, and returns the standard deviation of a 100 point rerandomization distribution for a statistic computed on each of 100 realizations of the allocation procedure. This is a fairly laborious undertaking that is embarassingly parallel, so the code given here uses the *multicore* package (Urbanek 2009) to dispatch the work over 12 cores. For the minimization

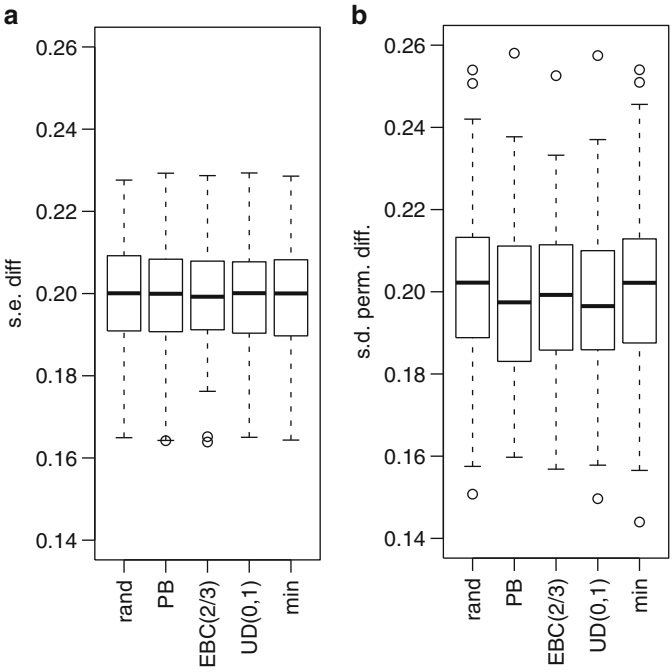

Fig. 2.4 Comparison of model-based standard error distribution (**a**) over 100 realizations of the sample cohort for five randomization algorithms, and permutation-based sampling standard deviations of the mean difference between arms for 100 permutations of each of the same realizations (**b**). Notation: rand denotes purely random allocation; PB denotes permuted blocks, fixed blocksize of 4; EBC(2/3) denoted Efron's biased coin with $p = 2/3$; UD(0,1) is Wei's urn design with $\alpha = 0$, $\beta = 1$, and min is Pocock–Simon minimization

procedure, reseeding the stratum-specific generator has limited impact on the sequence of covariate-based allocations, at best altering only the very first allocation directly. Therefore, we permute the arrival times of participants to the trial for this procedure.

Figure 2.5 shows that, relative to model-based standard error computation, there is a hint of variance reduction for several of the allocation schemes when rerandomization is employed in this context.

```
> getrrstats = function(rlist, ce) {
+     randomization(ce) = rlist
+     require(multicore)
+     NRERAND = 100
+     NREALIZ = 100
+     NPATS = 100
+     rerandstats = mclapply(1:NREALIZ, function(r) {
+       cat(r)
+       seedmat = matrix(round(1e+05 * runif(400), 0),
          nr = NRERAND)
```

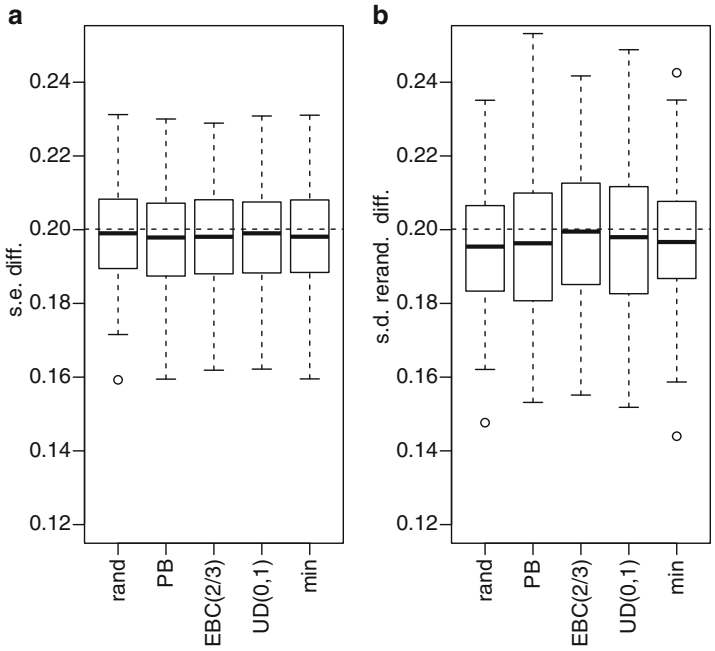

Fig. 2.5 Comparison of model-based standard error distribution (**a**) over 100 realizations of the sample cohort for five randomization algorithms, and rerandomization-based sampling standard deviations of the mean difference between arms for 500 rerandomizations of each of the same realizations (**b**). Rerandomization for the minimization algorithm employed permutation of the entry times for trial participants. Notation: rand denotes purely random allocation; PB denotes permuted blocks, fixed blocksize of 4; EBC(2/3) denoted Efron's biased coin with $p = 2/3$; UD(0,1) is Wei's urn design with $\alpha = 0$, $\beta = 1$, and min is Pocock–Simon minimization

```
+          rerandest = rep(NA, NRERAND)
+          for (i in 1:NRERAND) {
+              cat(i)
+              curtri = createTrial(ce, seed = seedmat
    [i, ])
+             shuff = function(x) x[sample(1:length(x),
                size = length(x),
+                   replace = FALSE)]
+              pdl = mallPDLISTS[[r]]
+              if (ce@randomization[[1]][[1]]@type
                == "Minimization")
+                  pdl = shuff(mallPDLISTS[[r]])
+              trts = lapply(1:NPATS, function(j)
                getTreatment(curtri,
+                  pdl[[j]]))
+              df = do.call(rbind, getEnrolleeInfo
                (curtri))
```

```
+              rerandest[i] = summary(lm(df$stdnor ~
                 df$alloc))$coef[2,
+                   "Estimate"]
+          }
+          sd(rerandest)
+      }, mc.cores = 12)
+ }
```

4 Discussion

Investigators planning randomized clinical experiments are fortunate to have access to the fruits of over 2 decades of statistical research on approaches to treatment allocation. The efforts that we have described in Sect. 1 confront universal requirements of protection from bias and maintenance of statistical efficiency in very different ways, and no broad consensus on best practices for the linked processes of treatment allocation and downstream analysis is currently available.

We have described a system that can be used to examine properties of different methods for allocating patients to treatments in clinical trials. Our intention is not to provide a fully functioning system that could be used for randomizing real patients, but rather to provide a system that is capable of addressing important questions in the design and analysis of clinical trials. The code is reasonably comprehensive and compact and could easily be extended by anyone familiar with the R programming language.

Figures 2.1–2.5 demonstrate some of the questions that can be addressed using this system. We are aware of no other systems that would allow for such explicit comparisons. More importantly, our system supports inference by rerandomization, and that, it seems, is essential for understanding other properties of the allocation schemes being used.

References

Begg CB, Iglewicz B (1980) A treatment allocation procedure for sequential clinical trials. Biometrics 36(1):81–90

Cox DR (1982) A remark on randomization in clinical trials. Utilitas Math 21A:245–252

Efron B (1971) Forcing a sequential experiment to be balanced. Biometrika 58:403–417

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. i. Introduction and design. Br J Cancer 34(6):585–612

Pocock SJ, Simon R (1975) Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 31(1):103–115

Rosenberger WF, Lachin JM (2002) Randomization in clinical trials: Theory and practice. Wiley, NY

Rosenberger WF, Sverdlov O (2008) Handling covariates in the design of clinical trials. Stat Sci 23(3):404–419

Taves D (1974) Minimization: A new method of assigning patients to treatment and control groups. Clin Pharmacol Ther 15(5):443–453. http://www.ncbi.nlm.nih.gov/pubmed/4597226

Urbanek S (2009) multicore: Parallel processing of R code on machines with multiple cores or CPUs, 2009. http://www.rforge.net/multicore/. R package version 0.1-3

Zelen M (1974) The randomization and stratification of patients to clinical trials. J Chron Dis 27(7–8):365–375

Chapter 3
Sequential Designs for Clinical Trials

KyungMann Kim

1 Introduction

Until the mid 1980s, when methods for group sequential design and analysis of clinical trials became available, most phase III randomized, controlled trials were planned with a fixed design in which the number of patients were predetermined, that is, fixed in advance, based on the design parameters. Patients enrolled serially over the period known as the accrual period, and after a fixed minimum follow-up after the last enrolled patient, a final analysis was performed. Notwithstanding such designs, however, because of ethical concern for the participating patients, unplanned interim analyses were often performed periodically on the accumulating data with consideration for possible early termination of the trial should there be sufficient evidence for beneficial or harmful effects. As will be noted later, these unplanned analyses have the potential to compromise the statistical integrity or fidelity of the trial.

Indeed, because of ethical and safety concern with experiments on humans, the conduct of clinical trials is often monitored at regular intervals to ensure the ongoing safety of participants and the validity and integrity of the data. This concern for the ethics of clinical trials and the safety of participants is particularly critical in phase III randomized, controlled trials. It is widely recognized that considerations should be given for possible early termination if continuation of the trial is unwarranted ethically when beneficial or adverse effects of a treatment are established with an acceptable or unacceptable risk/benefit ratio, respectively.

Sequential methods for analysis of accumulating data were developed in the 1920s (Dodge and Romig 1929) and mid 1940s (Wald 1947), establishing that

K. Kim (✉)
University of Wisconsin-Madison, 600 Highland Ave, K6/438 CSC,
Madison, WI 53792-4675, USA
e-mail: kmkim@biostat.wisc.edu

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7_3,

Table 3.1 Effects of repeated significance testing at a nominal 0.05 level

	Number of repeated significance tests										
	1	2	3	4	5	10	20	50	100	⋯	∞
Type I error	0.050	0.083	0.107	0.126	0.142	0.193	0.246	0.320	0.374	⋯	1.000

a savings in the required sample size is possible by applying hypothesis tests repeatedly without compromising the statistical integrity of the test by maintaining the type I error and type II error probabilities. Peter Armitage was perhaps the first to recognize the ethical relevance of sequential methods for clinical trials in a series of papers in later 1950s, culminating in his landmark textbook published in 1960 and updated later (Armitage 1975). Sequential methods are widely used lately in all phases of clinical trials, from phase I dose-finding trials, phase II efficacy screening trials, and to phase III effectiveness trials.

There are two approaches to sequential analysis of clinical trials, one based on repeated significance testing and the other based on the so-called sequential boundaries. The repeated significance testing approach adjusts significance levels to account for repeated or multiple tests in interim analyses. These methods have their origin in Armitage (1954) and were subsequently adapted by Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983). The approach based on sequential boundaries has its origin in the sequential probability ratio test (SPRT) of Wald (1947) and includes the triangular test initially developed by Anderson (1960) as a modification of the SPRT to reduce sample size and later adapted for interim analyses by Whitehead (1997). In this chapter, we will describe the designs based on the repeated significance testing approach.

If one were to apply the fixed sample significance test repeatedly, thus the terminology "repeated significance testing," the false positive rate, that is type I error probability, is known to become inflated beyond the desired level, ultimately converging to one. This phenomenon was described as *sampling to reach a foregone conclusion* by Anscombe (1954). Table 3.1 summarizes the effect of repeated significance testing for normal data at a nominal level of $\alpha = 0.05$ and shows the actual type I error probabilities achieved.

This can be explained by the law of the iterated logarithm and clearly indicates the need to adjust the significance levels or the critical values for the repeated significance tests. To maintain the overall significance level despite the repeated significance tests, one has to choose the critical values $c_1, \ldots, c_K$ of the repeated significance tests such that

$$\Pr(|Z_1| < c_1, \ldots, |Z_K| < c_K) = 1 - \alpha$$

under the null hypothesis, where each Z_k denotes the standardized test statistic based on all the data up to the kth interim analysis and K is the number of planned tests.

Given the critical values, the nominal significance level for each interim analysis is defined as

$$\Pr(|Z_k| > c_k) = \alpha_k,$$

and the exit probability is defined as

$$\Pr(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, |Z_k| > c_k) = \pi_k,$$

with the requirement $\sum_{k=1}^{K} \pi_k = \alpha$.

The first step in a group sequential design involves the determination of the critical values to maintain the type I error probability despite the interim analyses. These adjustments for the nominal significance levels also have an implication for the sample size requirements. After reviewing the sample estimation for a normal data problem, we will describe the sample size estimation in terms of the Fisher information and describe a very general framework for the estimation of the maximum information for group sequential design based on the efficient score test.

The rest of this chapter is organized as follows. Section 2 reviews sample size estimation for a fixed design based on normal data and generalizes it to estimation of the fixed information based on the efficient score test for normal, time to event, and dichotomous data. Section 3 reviews classical group sequential designs for normal data along with the associated computational issues and estimation of the maximum sample size. This section also introduces the notion of the inflation factor based on the relationship between the sample sizes for the fixed design and the group sequential design. Section 4 develops sample size estimation for information-based group sequential designs in terms of the maximum information based on the efficient score test and provides the theoretical justification for the use of the inflation factor in determining the maximum information given the information for the corresponding fixed design. Section 5 describes how information-based group sequential analysis is performed based on the type I error spending function, enabling the practical application of group sequential tests to a range of clinical trials. Section 6 presents an example of information-based group sequential design and analysis as applied to a phase III trial in cancer. Section 7 concludes this chapter with a discussion and some concluding remarks.

2 The Fixed Design

In this section, sample size estimation for a fixed design based on normal data will be reviewed, followed by generalization based on the efficient score test. The information-based approach will be described for normal, time to event, and dichotomous data for illustration.

2.1 Normal Data

Consider a randomized trial of two treatments with equal allocation fractions. Suppose the responses to treatment in the two groups are normally distributed with means μ_E and μ_C for the experimental and control treatment, respectively, with a common known variance σ^2, and one is interested in testing the null hypothesis of no treatment difference $H_0 : \mu_E = \mu_C$ against the alternative hypothesis $H_1 : \mu_E \neq \mu_C$, or equivalently

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta \neq 0,$$

where $\delta = \mu_E - \mu_C$. With the standardized test statistic based on a total of n patients under equal randomization

$$Z = \frac{\bar{X}_E - \bar{X}_C}{\sqrt{4\sigma^2/n}},$$

where $\bar{X}_E$ and $\bar{X}_C$ denote the sample means, one would reject H_0 if $|Z| > z_{\alpha/2}$, the upper $\alpha/2$ quantile of the standard normal distribution. The fixed design requires a total sample size of

$$n = \frac{4(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta_R^2}$$

to detect the difference δ_R with power $1 - \beta$ using a two-sided significance level α.

2.2 Efficient Score Test

In general, the comparison of treatment effects in a two-arm trial can be based on a test given by the efficient score statistic for the parameter of interest θ,

$$S \sim N(\theta\mathscr{I}, \mathscr{I}),$$

where θ represents a treatment difference and $\mathscr{I}$ the Fisher information about θ contained in S. In general, the efficient score statistic is derived from the likelihood function $L(\theta|X)$ of the parameter θ given the data X. For regular problems, the efficient score statistic S is the first order partial derivative of the log likelihood function $l(\theta|X) = \log L(\theta|X)$, that is,

$$S = \frac{\partial l(\theta|X)}{\partial\theta},$$

and the Fisher information $\mathscr{I}$ is the expectation of the negative second order partial derivative of the log likelihood function, that is,

$$\mathscr{I} = E\left\{-\frac{\partial^2 l(\theta|X)}{\partial\theta^2}\right\},$$

both evaluated at the true value of θ.

The null hypothesis of no difference is usually $H_0 : \theta = 0$, and the alternative hypothesis is $H_1 : \theta \neq 0$. A treatment comparison based on the approximate normal sampling distribution of S and critical value b, will have type I error probability α and power $1 - \beta$ to detect a specific treatment difference θ_R as long as

$$\Pr(S > b|\theta = 0) = \alpha/2$$

and

$$\Pr(S > b|\theta = \theta_R) = 1 - \beta.$$

These two requirements are met by

$$\mathscr{I} = \frac{(z_{\alpha/2} + z_\beta)^2}{\theta_R^2}$$

and

$$b = \frac{z_{\alpha/2}(z_{\alpha/2} + z_\beta)}{\theta_R} = z_{\alpha/2}\sqrt{\mathscr{I}}.$$

2.3 Fixed Information

With the normally distributed data, the efficient score statistic is given by

$$S = \frac{n(\bar{X}_E - \bar{X}_C)}{4\sigma}$$

with the standardized treatment difference, that is, the effect size

$$\theta = \frac{\mu_E - \mu_C}{\sigma}$$

and Fisher information

$$\mathscr{I} = \frac{n}{4},$$

where n is the total sample size with equal randomization between the two treatments. The total sample size n is thus

$$n = \frac{4(z_{\alpha/2} + z_\beta)^2}{\theta_R^2} = \frac{4(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta_R^2},$$

the right-hand side being the familiar sample size formula given in Sect. 2.1. Note that the test statistic Z given in Sect. 2.1 is simply a standardized form of the efficient score statistic

$$Z = \sqrt{\frac{n}{4}}\frac{\bar{X}_E - \bar{X}_C}{\sigma} = S/\sqrt{\mathscr{I}}.$$

With time to event data subject to right censoring, the comparison of the two survival distributions is often based on the logrank test. Under the proportional hazards assumption

$$\lambda(t|z) = \lambda_0(t)\exp(\theta z),$$

with $z = 0, 1$, respectively, denoting the control or experimental treatment, the null hypothesis of no difference is $H_0 : \theta = 0$, where $\theta = \log \Delta$ is the constant log hazard ratio. Under the proportional hazards model, the logrank test statistic is equivalent to the efficient score statistic for the log hazard ratio θ with Fisher information $\mathscr{I} = d/4$, where d is the total number of events to be observed. Therefore, the total number of events necessary to detect the hazard ratio Δ_R between the two treatments with power $1 - \beta$ at a two-sided significance level α is given by

$$d = \frac{4(z_{\alpha/2} + z_\beta)^2}{\log^2 \Delta_R}.$$

There are two approaches to the design of randomized trials with time to event data: calendar time-driven duration trials versus event-driven information trials (Lan and DeMets 1989). They are different in the way the end of a trial is defined. With the duration trial, patients are followed over a fixed duration of follow-up for a specified number of patients (a sample size in a traditional sense) enrolled during a fixed duration of accrual (Lan and DeMets 1989; Lan and Lachin 1990; Kim and Tsiatis 1990; Kim 1992; Kim et al. 1995). With the information trial, the trial is concluded when a prespecified number of events have been observed (Lan and DeMets 1989; Kim 1995; Kim et al. 1995).

With dichotomous data, comparison of the probability of "success" can again be based on the efficient score statistic for the log odds ratio

$$\theta = \log \frac{p_E(1 - p_C)}{p_C(1 - p_E)},$$

where p_E and p_C are the success probabilities for the experimental and control treatment, respectively. The efficient score statistic for θ is

$$S = \frac{n(\hat{p}_E - \hat{p}_C)}{4},$$

where n is the total sample size under equal randomization and $\hat{p}_E$ and $\hat{p}_C$ are the estimated probabilities of success. In this setting, the Fisher information is

$$\mathscr{I} = \frac{n\bar{p}(1 - \bar{p})}{4},$$

where $\bar{p} = (p_E + p_C)/2$. Therefore, the total sample size necessary to detect the log odds ratio θ_R between the two treatments with power $1-\beta$ under the null hypothesis of no difference $H_0 : \theta = 0$ at a two-sided significance level α is given by

$$n = \frac{4(z_{\alpha/2} + z_\beta)^2}{\theta_R^2 \bar{p}(1-\bar{p})}.$$

This formula gives a smaller sample size estimate than the usual sample size formula for a two-sample binomial problem given by either

$$n = \frac{4(z_{\alpha/2} + z_\beta)^2 \bar{p}(1-\bar{p})}{(p_E - p_C)^2}$$

or

$$n = \frac{2\{z_{\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_\beta \sqrt{p_E(1-p_E) + p_C(1-p_C)}\}^2}{(p_E - p_C)^2},$$

both of which are based on a normal approximation.

3 Classical Group Sequential Design

In this section, classical group sequential designs for normal data will be reviewed along with the associated computational issues and estimation of the maximum sample size. We also discuss the use of an inflation factor based on the relationship between the sample sizes for the fixed design and the group sequential design.

Consider the normal data setting in Sect. 2. Instead of performing one analysis after n subjects, suppose one plans to perform interim analyses of the accumulating data up to K times after every n subjects. Then, for the kth interim analysis, one would consider the partial sum

$$S_k = \sum_{j=1}^{k} Y_j \sim N(\delta^\star k, k),$$

where

$$Y_j = \frac{\bar{X}_{Ej} - \bar{X}_{Cj}}{\sqrt{4\sigma^2/n}} \sim N(\delta^\star, 1)$$

with $\delta^\star = \delta/\sqrt{4\sigma^2/n}$ and $\bar{X}_{Ej}$ and $\bar{X}_{Cj}$, the sample means of observations accumulated between the $(j-1)$st and the jth interim analyses. Equivalently, one could compute the standardized test statistic

$$Z_k = \frac{\sum_{j=1}^{k}(\bar{X}_{Ej} - \bar{X}_{Cj})}{\sqrt{k(4\sigma^2/n)}} = \frac{S_k}{\sqrt{k}} \sim N(\delta^\star\sqrt{k}, 1)$$

Table 3.2 Critical and boundary values for two-sided group sequential designs with $\alpha = 0.05$ and $K = 5$

	Pocock		O'Brien-Fleming	
k	Critical	Boundary	Critical	Boundary
1	2.41	2.41	4.56	4.56
2	2.41	3.41	3.23	4.56
3	2.41	4.18	2.63	4.56
4	2.41	4.83	2.28	4.56
5	2.41	5.40	2.04	4.56

Table 3.3 Nominal significance levels and exit probabilities for two-sided group sequential designs with $\alpha = 0.05$ and $K = 5$

	Pocock		O'Brien-Fleming	
k	Nominal	Exit	Nominal	Exit
1	0.0158	0.0158	0.00000504	0.00000504
2	0.0158	0.0117	0.00125	0.00125
3	0.0158	0.0090	0.00843	0.00765
4	0.0158	0.0074	0.0225	0.0167
5	0.0158	0.0061	0.0413	0.0244

and decide whether to stop or to continue to the next interim analysis, depending on whether the partial sum S_k exceeds the boundary value b_k or the test statistic Z_k exceeds the critical value c_k.

Pocock (1977) suggested using a constant critical value at each interim analysis and rejecting H_0 when

$$|Z_k| \geq c_k \equiv c_{\mathrm{P}} \text{ or equivalently } |S_k| \geq b_k = c_{\mathrm{P}}\sqrt{k},$$

whereas O'Brien and Fleming (1979) suggested using a constant boundary value at each interim analysis and rejecting H_0 when

$$|S_k| \geq b_k \equiv c_{\mathrm{O}}\sqrt{K} \text{ or equivalently } |Z_k| \geq c_k = c_{\mathrm{O}}\sqrt{K/k}.$$

With the overall significance level $\alpha = 0.05$ and the number of repeated significance tests $K = 5$, $c_{\mathrm{P}} = 2.41$ for Pocock group sequential test and $c_{\mathrm{O}} = 2.04$ for O'Brien-Fleming group sequential test. Table 3.2 gives the critical and boundary values for the two-sided Pocock and O'Brien-Fleming group sequential designs when $\alpha = 0.05$ and $K = 5$, while Table 3.3 gives the nominal significance levels $\alpha_k = 2\{1 - \Phi(c_k)\}$, where Φ is the standard normal distribution function and π_k are the exit probabilities.

Figure 3.1 shows the critical values c_k, $k = 1 : K$, for the two-sided group sequential tests by Pocock (1977) and by O'Brien and Fleming (1979) when $\alpha = 0.05$ and $K = 5$, while Fig. 3.2 shows the boundary values b_k, $k = 1 : K$. Note that it is difficult to reject H_0 early with O'Brien-Fleming group sequential tests because of larger critical values at earlier interim analyses, with the final test becoming similar to a fixed sample test. It is easier to reject H_0 early with Pocock group sequential tests, however, the last critical value for Pocock group sequential tests can be much larger than the critical value for the fixed sample test.

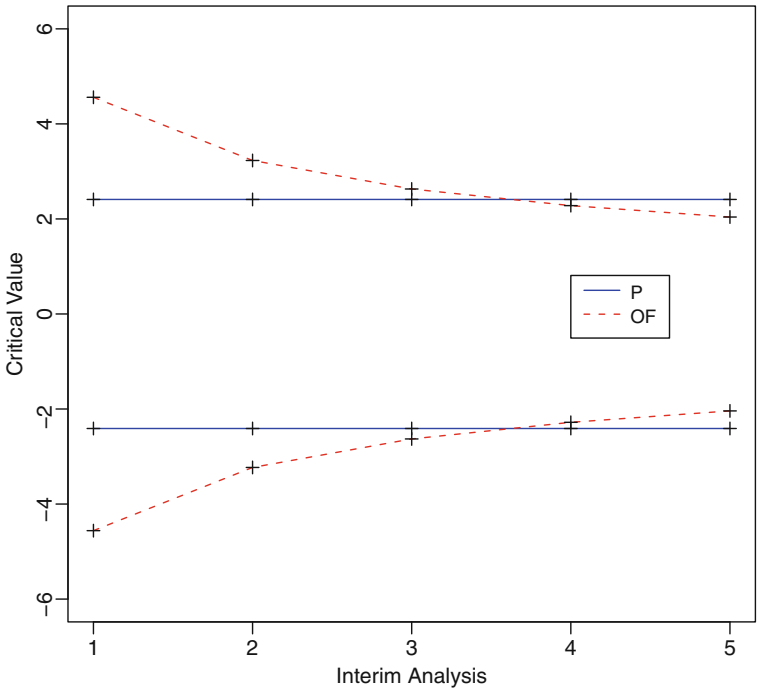

Fig. 3.1 Critical values for two-sided group sequential designs with $\alpha = 0.05$ and $K = 5$ (P for Pocock and OF for O'Brien-Fleming)

3.1 Computational Issues

The computations required for Pocock and O'Brien-Fleming group sequential designs can be performed using the recursive numerical integration procedure by Armitage et al. (1969) under the null hypothesis, and by McPherson and Armitage (1971) under the alternative hypothesis as described below. As noted in Sect. 2, group sequential designs for comparison of normal outcomes between two treatments in a randomized trial can be simplified to a sequential design for a one-sample normal data.

Let, Y_k denote the standardized test statistic for the kth interim analysis for the classical group sequential test

$$Y_k = \frac{\bar{X}_{Ek} - \bar{X}_{Ck}}{\sqrt{4\sigma^2/n}} \sim N(\delta^\star, 1)$$

introduced earlier, so that $Y_1, \ldots, Y_K$ are independently and identically distributed with mean $\delta^\star$ and variance one. Note that $Y_1, \ldots, Y_K$ are independent since each Y_k is based on independent observations accumulated between the $(k-1)$st and the kth

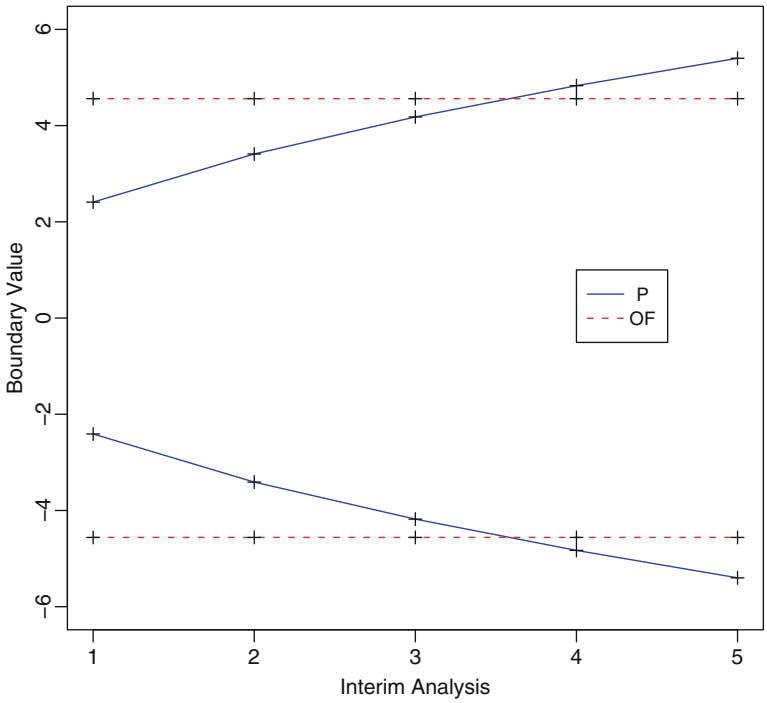

Fig. 3.2 Boundary values for two-sided group sequential designs with $\alpha = 0.05$ and $K = 5$ (P for Pocock and OF for O'Brien-Fleming)

interim analyses, and identically distributed because of equal increments in sample size between successive interim analyses. Let $S_k = \sum_{j=1}^{k} Y_j$ denote the partial sum, and $Z_k = S_k/\sqrt{k}$ its standardized statistic as before.

Denote by f_k the probability density function of S_k in the sequential sampling. Because of the independent increment structure in the partial sum S_k, the probability density for S_k can be determined recursively using the convolution of the distributions for S_{k-1} and Y_k,

$$f_k(s) = \int_{-b_{k-1}}^{b_{k-1}} f_{k-1}(u)\phi(s-u)\mathrm{d}u,$$

where f_1 is the standard normal density function ϕ. Since the probability of stopping at or before the kth interim analysis is

$$1 - \Pr(|S_1| < b_1, \ldots, |S_k| < b_k) = 1 - \int_{-b_k}^{b_k} f_k(u)\mathrm{d}u,$$

the type I error probability is determined by

$$\alpha = 1 - \int_{-b_K}^{b_K} f_K(u)\mathrm{d}u$$

under the null hypothesis. Alternatively, with the prespecified exit probability π_k of stopping exactly at the kth interim analysis given by

$$\begin{aligned}\pi_k &= \Pr(|S_1| < b_1, \ldots, |S_{k-1}| < b_{k-1}, |S_k| \geq b_k) \\ &= \int_{-b_{k-1}}^{b_{k-1}} f_{k-1}(u)\{1 - \Phi(b_k - u) + \Phi(-b_k - u)\}\mathrm{d}u,\end{aligned}$$

the overall type I error probability becomes $\alpha = \pi_1 + \cdots + \pi_K$, as described by Slud and Wei (1982).

3.2 Maximum Sample Size and Inflation Factor

As noted earlier, given the overall significance level α and the number of repeated significance tests K, one can determine the critical values or equivalently the boundary values for the chosen classical group sequential design. With the Pocock group sequential design, one would determine the constant critical values $c_k \equiv c_{\mathrm{P}}(\alpha, K)$ or alternatively the boundary values $b_k = c_{\mathrm{P}}(\alpha, K)\sqrt{k}$, $k = 1 : K$. With O'Brien-Fleming group sequential design, one would determine the critical values $c_k = c_{\mathrm{O}}(\alpha, K)\sqrt{K/k}$ or the constant boundary values $b_k \equiv c_{\mathrm{O}}(\alpha, K)\sqrt{K}$. This is accomplished by solving the equation

$$1 - \alpha = \int_{-b_K}^{b_K} f_K(u)\mathrm{d}u$$

for either $c_{\mathrm{P}}(\alpha, K)$ or $c_{\mathrm{O}}(\alpha, K)$ using the computational procedure by Armitage et al. (1969).

To design a randomized trial as a classical group sequential trial, one needs to determine the number of patients n for two treatments under equal randomization between successive interim analyses, which in turn determines the maximum sample size for the chosen sequential design. Given the critical or the boundary values, the power $1 - \beta$ to detect the difference in mean between two treatments δ_{R} can be determined as

$$1 - \beta = \Pr(|Z_k| < c_k, k = 1 : K).$$

Conversely, given the desired power $1 - \beta$ to detect the treatment difference δ_{R}, one can determine the value of $\delta^{\star}(\alpha, K, \beta)$ by solving the above equation using

the recursive numerical integration procedure by McPherson and Armitage (1971). Since $\delta^\star(\alpha,K,\beta)=\delta_R/\sqrt{4\sigma^2/n}$, one then solves for n as

$$n=\frac{4\{\delta^\star(\alpha,K,\beta)\}^2\sigma^2}{\delta_R^2}.$$

The maximum sample size for the group sequential design is then given by

$$nK=\frac{4\{\delta^\star(\alpha,K,\beta)\sqrt{K}\}^2\sigma^2}{\delta_R^2}.$$

Since the sample size for the fixed design with $K=1$ is

$$n=\frac{4(z_{\alpha/2}+z_\beta)^2\sigma^2}{\delta_R^2},$$

$$\delta^\star(\alpha,1,\beta)=z_{\alpha/2}+z_\beta.$$

For a classical group sequential design with $K>1$, the maximum sample size nK as derived above is

$$\begin{aligned}nK&=4\{\delta^\star(\alpha,K,\beta)\sqrt{K}\}^2\left(\frac{\sigma}{\delta_R}\right)^2\\&=4(z_{\alpha/2}+z_\beta)^2\left(\frac{\sigma}{\delta_R}\right)^2\times\frac{\{\delta^\star(\alpha,K,\beta)\sqrt{K}\}^2}{(z_{\alpha/2}+z_\beta)^2},\end{aligned}$$

a constant multiple of the corresponding sample size for the fixed design. This constant multiple

$$\mathscr{F}(\alpha,K,\beta)=\frac{\{\delta^\star(\alpha,K,\beta)\sqrt{K}\}^2}{(z_{\alpha/2}+z_\beta)^2}$$

is referred to as the inflation factor. Note that with $K=1$, that is, with the fixed design, it reduces to one as $\delta^\star(\alpha,1,\beta)=z_{\alpha/2}+z_\beta$ as noted above, leading to the sample size for the fixed design. Kim and DeMets (1987,1992) provides the square root of the numerators for the inflation factors for various group sequential designs.

Table 3.4 reproduces from Kim et al. (2003) the inflation factors for the group sequential designs with the number of planned repeated significance tests, $K=2,3,4,5$, for both Pocock and O'Brien-Fleming designs, with power, $1-\beta=0.8,0.90,0.95$, at a two-sided test with significance levels $\alpha=0.05,0.01$. Note that Pocock group sequential designs require an up-front commitment of 8–23% more in sample size than the corresponding fixed design, while O'Brien-Fleming group sequential designs require a minimal increase from the fixed design.

Under certain conditions, the inflation factor simplifies the sample size estimation for many types of outcome data. If the test statistic has equal increments between

Table 3.4 Inflation factors for two-sided group sequential designs (P for Pocock and OF for O'Brien-Fleming

		K							
		2		3		4		5	
α	$1-\beta$	P	OF	P	OF	P	OF	P	OF
0.05	0.80	1.11	1.01	1.17	1.02	1.20	1.02	1.23	1.03
	0.90	1.10	1.01	1.15	1.02	1.18	1.02	1.21	1.03
	0.95	1.09	1.01	1.14	10.2	1.17	1.02	1.19	1.02
0.01	0.80	1.09	1.00	1.14	1.01	1.17	1.01	1.19	1.02
	0.90	1.08	1.00	1.12	1.01	1.15	1.01	1.17	1.01
	0.95	1.08	1.00	1.12	1.01	1.14	1.01	1.16	1.01

successive interim analyses and an independent increment structure, the maximum sample size necessary for the group sequential design can easily be determined simply by multiplying the sample size for the corresponding fixed design by the inflation factor. This simple calculation follows from the analogy between the partial sum process for the normal data and the efficient score statistics with fixed drift and variance which is Fisher information for the fixed drift with equal increments. This will be discussed in detail in the next section.

4 Information-Based Group Sequential Design

In this section, sample size estimation for the so-called information-based group sequential designs is developed in terms of the maximum information based on the efficient score test, followed by the theoretical justification for the use of the inflation factor in determining the maximum information for the information-based group sequential design given the information for the corresponding fixed design.

Suppose only one analysis is to be performed at calendar time T, as in a fixed design, based on the efficient score test

$$S(T) \sim N(\theta \mathscr{I}(T), \mathscr{I}(T)),$$

with drift parameter θ representing the parameter of interest and Fisher information $\mathscr{I}(T)$ or, equivalently, on the Wald test

$$Z(T) = \frac{\hat{\theta}(T)}{se\{\hat{\theta}(T)\}} \sim N(\theta\sqrt{\mathscr{I}(T)}, 1).$$

In the Wald test, $\hat{\theta}(T)$ denotes the maximum likelihood estimate of θ and $se\{\hat{\theta}(T)\}$ the estimated standard error of $\hat{\theta}(T)$, which satisfies

$$\frac{se^{-2}\{\hat{\theta}(T)\}}{\mathscr{I}(T)} \to 1$$

in probability. In testing $H_0 : \theta = 0$ at a significance level α against $H_A : \theta \neq 0$ with power $1 - \beta$ to detect the treatment effect $\theta = \theta_R$, a normal approximation can be used to show,

$$\mathscr{I}(T) = \frac{(z_{\alpha/2} + z_\beta)^2}{\theta_R^2},$$

the fixed information as derived in Sect. 2.1.

For a general class of parametric and semiparametric models, a Wald test computed at time t satisfies

$$Z(t) = \frac{\hat{\theta}(t)}{se\{\hat{\theta}(t)\}} \sim N(\theta\sqrt{\mathscr{I}(t)}, 1).$$

Equivalently,

$$S_W(t) = \sqrt{\mathscr{I}(t)}Z(t) \sim N(\theta\mathscr{I}(t), \mathscr{I}(t)),$$

is a Brownian motion process with fixed drift θ and variance given by the Fisher information $\mathscr{I}(t)$, with

$$\frac{se^{-2}\{\hat{\theta}(t)\}}{\mathscr{I}(t)} \to 1$$

in probability. This test statistics is thus equivalent to the efficient score statistic

$$S(t) \sim N(\theta\mathscr{I}(t), \mathscr{I}(t)).$$

Therefore, the statistical information for the parameter of interest θ is the precision of the efficient estimator of θ having the smallest variance which can be approximated by $se^{-2}\{\hat{\theta}(t)\}$.

Suppose one plans to perform interim analyses and the final analysis in a randomized, controlled trial at calendar times $t_1, \ldots, t_K$. Then for the kth interim analysis at time t_k, one may compute either Wald test

$$Z(t_k) = \frac{\hat{\theta}(t_k)}{se\{\hat{\theta}(t_k)\}}$$

or the efficient score statistic $S(t_k)$ and terminate the trial early and reject $H_0 : \theta = 0$ if $|Z(t_k)| \geq c_k$ or $|S(t_k)| > b_k$, where c_k and b_k are the critical and boundary values for the group sequential test, respectively, as in Sect. 3.

Typically, the joint distribution of the group sequential test statistics is multivariate normal or at least asymptotically so, and subsequently group sequential methods require multivariate numerical integration. If the increments in the test statistic between successive interim analyses are independent, however, the multivariate numerical integration reduces to univariate numerical integrations based on a simple recursion involving a convolution of two random variables as in Armitage et al. (1969) and McPherson and Armitage (1971). As shown by Scharfstein et al.

(1997), any efficient-based test statistic that achieves a semi-parametric efficiency bound – almost all test statistics used in practice – calculated at calendar times $t_1, \ldots, t_K$ behaves like a standardized partial sum of independent normal variables and thus has independent increments between successive interim analyses. More specifically, under a general setting, the joint distribution of the vector of sequential test statistics $\{S(t_1), \ldots, S(t_K)\}$ is asymptotically normal with mean vector

$$\{\theta \mathscr{I}(t_1), \ldots, \theta \mathscr{I}(t_K)\}$$

and the covariance structure of an independent increment process:

$$\mathrm{Var}\{S(t_k)\} = \mathscr{I}(t_k)$$

and, for $k \leq l$,

$$\mathrm{Cov}[S(t_k)\{S(t_l) - S(t_k)\}] = 0.$$

This property of the sequential test statistics over time remains true even under contiguous alternatives. Jennison and Turnbull (1997) showed the same for a general class of regression models as well.

Given K, assume at least tentatively that the interim analyses will be performed at calendar times $t_1, \ldots, t_K$ after equal increments of information, that is,

$$\mathscr{I}(t_k) = \frac{k \mathscr{I}_{\max}}{K},$$

where $\mathscr{I}_{\max} = \mathscr{I}(t_K)$. Then with the information fraction $\tau_k = \mathscr{I}(t_k)/\mathscr{I}(t_K) = k/K$ and $\eta = \theta\sqrt{\mathscr{I}(t_K)} = \delta^\star\sqrt{K}$ the standardized efficient score statistic is

$$\frac{S(t_k)}{\sqrt{\mathscr{I}(t_K)}} \sim N(\eta\tau_k, \tau_k).$$

The statistic is equivalent to

$$\frac{S_k}{\sqrt{K}} \sim N(\eta\tau_k, \tau_k)$$

for the classical group sequential test for normal data described in Sect. 3. Because of this equivalence in distribution, once the information for a fixed design is known, the maximum information for the information-based group sequential design is determined simply as

$$\mathscr{I}_{\max} = \frac{(z_{\alpha/2} + z_\beta)^2}{\theta_R^2} \times \mathscr{F}(\alpha, K, \beta),$$

where $\mathscr{F}(\alpha, K, \beta)$ is the inflation factor derived in Sect. 3.2.

Note that the result above works only when there is an independent increment structure in the test statistics and there are equal increments between successive interim analyses.

5 Information-Based Group Sequential Analysis

In this section, the information-based group sequential analysis based on the type I error spending function is described. This approach allows the practical application of group sequential tests in the common situation in which the assumptions required for the classical group sequential design are not feasible; interim analyses are often inserted, adding to the prespecified number of analyses, and the increments of statistical information between successive analyses are rarely equal.

As noted in Sect. 4, the efficient score statistic has an independent increment structure. The information-based group sequential design described in Sect. 4 requires two other conditions, one on equal increments of information between successive interim analyses and the other on the number of planned repeated significance tests. In many long-term chronic disease clinical trials, the primary outcome of interest is either time to event such as death with right censoring or repeated measures taken at successive follow-up visits with potential missing data. Typically, patients enter clinical trials serially in a way known as staggered entry. As a consequence, it is generally unclear whether the requirement regarding the independent increments between successive interim analyses is to be met, and there is no guarantee that the interim analyses will be performed after equal increments of statistical information. Also, it is likely that the actual number of repeated significance tests performed would deviate from the prespecified number of repeated significance tests. Therefore, flexible methods are required in group sequential analysis.

To address these problems, Slud and Wei (1982) suggest that one prespecify exit probabilities for group sequential tests as

$$\pi_k = \Pr(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, |Z_k| \geq c_k),$$

such that $\sum_{k=1}^{K} \pi_k = \alpha$, the overall significance level. The problem with this approach is the choice of the exit probabilities is arbitrary. Recognizing that both Pocock and O'Brien-Fleming group sequential designs can be interpreted from the exit probability perspective as indicated in Table 3.5 and in Fig. 3.3, Lan and DeMets (1983) instead suggested specifying a type I error spending function, to allow interim analyses at arbitrary calendar times in unequal increments of statistical information or even sporadically, doing fewer than or more than the planned K repeated significance tests.

The type I error spending function is defined as a monotonically increasing function, $\alpha^{\star}(\tau)$, of the information fraction τ, $0 \leq \tau \leq 1$, with $\alpha^{\star}(0) = 0$ and

Table 3.5 Cumulative exit probabilities for two-sided group sequential designs with $\alpha = 0.05$ and $K = 5$

k	Pocock	O'Brien-Fleming
1	0.0158	0.00000504
2	0.0275	0.00126
3	0.0365	0.00891
4	0.0439	0.0256
5	0.0500	0.0500

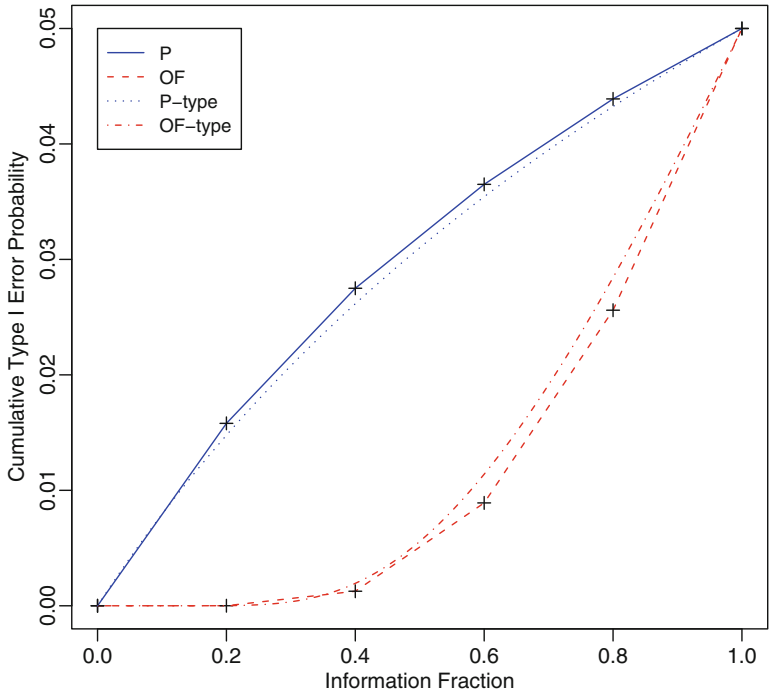

Fig. 3.3 Cumulative exit probabilities and error spending functions (P for Pocock and OF for O'Brien-Fleming; P-type for $\alpha_2^\star(\tau)$ and OF-type for $\alpha_1^\star(\tau)$)

$\alpha^\star(1) = \alpha$. Note that $\alpha^\star(\tau_k) - \alpha^\star(\tau_{k-1}) = \pi_k$ specifies the exit probability above. To perform interim analyses using the type I error spending function, the information fractions τ_k have to be estimated. In general, the information fraction is defined as $\tau_k = \mathscr{I}(t_k)/\mathscr{I}_{\max}$ for the efficient score test as in Sect. 4. As a special case, τ_k for both Pocock and O'Brien-Fleming group sequential designs is k/K, again as in Sect. 4. In general, there is no direct relationship between the information fraction and the calendar time of interim analysis. In fact, the information fraction is the same as the time of the discretized Brownian motion process on a unit interval. Subsequently, Kim and DeMets (1987) developed a general procedure for design of such group sequential trials.

For the Pocock and O'Brien-Fleming group sequential designs, Lan and DeMets (1983) proposed, respectively, the following type I error spending functions,

$$\alpha_2^\star(\tau) = \alpha \log\{1 + (e-1)\tau\}$$

and

$$\alpha_1^\star(\tau) = 2\{1 - \Phi(z_{\alpha/2}/\sqrt{\tau})\}$$

for one-sided tests. Figure 3.3 shows the plot of these two spending functions as well as the cumulative exit probabilities for the Pocock and O'Brien-Fleming group sequential designs. Note how close the graphs for $\alpha_2^\star(\tau)$ and $\alpha_1^\star(\tau)$ are, respectively, to the plots of the cumulative exit probabilities for Pocock and O'Brien-Fleming group sequential designs with $\alpha = 0.05$ and $K = 5$.

Now consider a group sequential trial. According to the information-based group sequential design with the overall significance level, α, the number of planned repeated significance tests, K, and the power, $1-\beta$, to detect a treatment difference θ_R, one would determine the maximum information $\mathscr{I}_{\max}$ and decide on the type I error spending function $\alpha^\star$ that has the characteristics of the chosen group sequential design. Instead of analyzing the data only once at the end of the trial, one may compute the test statistics at calendar times t_k, $k = 1:K$, of interim analyses. The strategy for stopping the trial early is to reject the null hypothesis H_0 the first time the test statistic is sufficiently large, that is,

$$|Z(t_k)| \geq c_k \text{ or } |S(t_k)| \geq b_k$$

for Wald test $Z(t_k)$ or the efficient score statistic $S(t_k)$, respectively. The critical value c_k or the boundary value b_k would be determined to satisfy

$$\begin{aligned}\pi_k = \alpha^\star(\tau_k) - \alpha^\star(\tau_{k-1}) &= \Pr(|Z(t_1)| < c_1, \ldots, |Z(t_{k-1})| < c_{k-1}, |Z(t_k)| \geq c_k) \\ &= \Pr(|S(t_1)| < b_1, \ldots, |S(t_{k-1})| < b_{k-1}, |S(t_k)| \geq b_k).\end{aligned}$$

Here, τ_k is estimated as the observed Fisher information for the efficient score statistic divided by $\mathscr{I}_{\max}$ or equivalently as

$$\tau_k = \frac{se^{-2}\{\hat{\theta}(t_k)\}}{\mathscr{I}_{\max}}.$$

A computational procedure very similar to that of Armitage et al. (1969) can be used for determining the critical or boundary values. The only difference is that the variances of the accumulating random variables in the partial sum are not the same. Because of the independent increment structure, recursive univariate integration can be used in place of a full multivariate normal integration. The computational procedures for the type I error spending function approach have been implemented in a suite of publicly available software as described in Reboussin et al. (2000).

6 An Example: CALGB 8433

In this section, an example of an information-based group sequential design and analysis in a phase III randomized, controlled trial in cancer is described briefly, illustrating the estimation of the maximum information based on the inflation fraction for group sequential design and the type I error spending function approach.

The Cancer and Leukemia Group B (CALGB) trial 8433 (Dillman et al. 1990) was designed to compare the standard treatment of radiotherapy, delivered over 6 weeks to the original tumor volume and involved regional lymph nodes, to an experimental treatment with 5 weeks of cisplatin plus vinblastine prior to the same radiotherapy in patients with stage III NSCLC. The rationale for the experimental treatment was that up-front systemic chemotherapy might lead to initial tumor shrinkage which would then improve the local control provided by radiotherapy and potentially eliminate micrometastatic disease. The primary objective of the trial was to compare overall survival on the experimental treatment to that on the standard treatment. This trial was terminated early before reaching its accrual goal due to appearance of a significant treatment difference between two treatments. Although this trial was not originally designed as a group sequential trial, group sequential analyses were implemented after patient accrual began and subsequently used as a basis for the decision for early termination. The group sequential design and analysis for this trial has been reported as a case study in a book chapter by Propert and Kim (1992).

Under the proportional hazards assumption, Cox model for failure time data is given by

$$\lambda(s|z) = \lambda_0(s)\exp(\theta z),$$

with the baseline hazard $\lambda_0(s)$ at time s, a treatment indicator z, and the log hazard ratio θ between the two treatments. The null hypothesis is $H_0 : \theta = 0$, that is, the log hazard ratio is zero, indicating that the hazard of death on the two treatments are the same, and the alternative hypothesis is $H_1 : \theta \neq 0$, indicating otherwise. They are equivalent to $H_0 : \Delta = 1$ and $H_1 : \Delta \neq 1$, respectively, with the constant hazard ratio $\Delta = \exp(\theta)$ between the two treatments.

The sample size was obtained to achieve power $1 - \beta = 0.8$ to detect a constant hazard ratio $\Delta = 1.5$, based on using a logrank test at a two-sided significance level $\alpha = 0.05$. This hazard ratio represents a 50% improvement in median survival with the experimental treatment over the standard treatment. By the fixed information formula for time to event data as described in Sect. 2.3

$$d = \frac{4(z_{\alpha/2} + z_\beta)^2}{\log^2 \Delta} = \frac{4(1.96 + 0.84)^2}{\log^2 1.5} = 190,$$

a total of 190 deaths on the two treatments were required to be observed during the trial. The sample size in the traditional sense was then determined by assuming that the analysis of the trial would take place after 80% of the patients would have died,

Table 3.6 Summary of group sequential analysis in CALGB 8433

k	t_k	τ_k	Logrank p-value	α_k	Decision
1	Sep 1985	0.05	n/a	0.0013	Keep open
2	Mar 1986	0.08	0.021	0.0013	Keep open
3	Aug 1986	0.18	0.0071	0.0013	Keep open
4	Oct 1986	0.22	0.0015	0.0013	Keep open
5	Mar 1987	0.29	0.0015[a]	0.0013	Close

[a] The p-value (0.0008) from the Cox model was used instead in the decision for early termination of the trial

so that 190/0.8 or approximately 240 patients were required to be enrolled into the trial, indicating that this trial was designed as a maximum information trial.

Based on previous experience in the same patient population, about 60 patients were expected to be accrued per year. Therefore, the trial was expected to require about 4 years of accrual with possibly 6 months to 1 year of additional follow-up to obtain 190 deaths. When this trial was designed initially in 1983, it was planned as a fixed design, and there was no provision for interim analysis and early stopping. The CALGB policies for interim analysis for possible early stopping were amended in 1986, coinciding with the emergence of treatment differences in CALGB 8433.

At the time of the first interim analysis, it was decided that an O'Brien-Fleming type I error spending function, $\alpha_1^\star(\tau)$, described in Sect. 5 would be used for formal sequential tests to take advantage of its flexible nature. However, to avoid extreme conservatism during the early phase of the trial with limited statistical information, $\alpha_1^\star(\tau)$ would be used with truncation at $c_k = 3$ with $\alpha_k = 0.0013$. Since a formal interim analysis was going to be performed, it was decided that the final analysis would be performed with more information than would be necessary for the fixed design to maintain the same power. This change in the trial from the fixed design to the sequential design at power $1 - \beta = 0.8$ needs no more than 5% increase in information from that for the corresponding fixed design. Therefore, the maximum number of deaths to be observed were inflated to $190 \times 1.05 = 200$.

Table 3.6 summarizes the group sequential analysis of the trial leading to early termination. Patient accrual to this trial began in May of 1984. Initial accrual was slow, and the first interim analysis was not performed until the fall of 1985. The sample size at this time (10 deaths in 50 patients) was too small to allow any meaningful comparisons of overall survival by treatment. At the next regularly scheduled interim analysis in the spring of 1986, the logrank test of the difference in overall survival between the two treatments gave an observed significance of $p = 0.021$. At the time, the median follow-up was only 6 months.

In August of 1986, with $\tau_3 = 0.18$, the logrank p-value was 0.0071. As a result of this interim analysis, a decision was made to perform an additional interim analysis 2 months later, for which a strong effort was made by the group to collect the most up-to-date follow-up information available for patients on the trial. By October of 1986, an additional eight failures had been observed, increasing the information fraction from 0.18 to $\tau_4 = 0.22$. With an increase in information fraction of 0.04,

the observed significance for the test of the difference in survival had decreased to $p = 0.0015$, which almost touched the truncated O'Brien-Fleming boundary significance level of 0.0013. Despite the close proximity to the boundary, it was decided that there was insufficient evidence to recommend closure of the trial and that the next interim analysis would be performed as scheduled in March of 1987.

The March 1987 interim analysis led to the decision to close the trial to further accrual. At that time, 163 out of the projected 240 patients had been accrued, and follow-up data were available from 105 patients with the median follow-up of 8 months. The logrank p-value for this comparison was 0.0015 with a total of 56 failures representing an information fraction of $\tau_5 = 0.29$. In response to concern that the observed treatment differences might be attributable to the differences in the two treatment groups with respect to prognostic factors such as age or substage of disease, a comprehensive analysis including use of the Cox (1972) proportional hazards model was undertaken. The prognostic factors examined were comparable between the two treatment arms. The results of the Cox analysis, controlling for various prognostic factors known for NSCLC, gave an observed significance of $p = 0.0008$. This analysis reaffirmed that the observed difference represented a real treatment effect. This adjusted p-value had crossed the truncated O'Brien-Fleming boundary, and it was unanimously voted to close the trial.

7 Discussion

The efficient score test for different types of outcome data in comparative clinical trials provides a unified framework in which to develop a most general form of group sequential designs. Recognizing the distributional similarity between the two-sample normal test for normal outcome data in classical group sequential designs and the efficient score test or equivalently Wald test, sample size estimation or estimation of maximum information can be simplified based on the notion of the inflation factor. Under the assumptions of the independent and equal increments of statistical information between successive interim analyses, the standard method for group sequential designs by Pocock (1977) and O'Brien and Fleming (1979) for normal data can be directly applied to information-based group sequential design and subsequent information-based group sequential analysis based on the type I error spending function by Lan and DeMets (1983).

Once a fixed design is known for any setting, an information-based group sequential design is automatic based on the efficient score test and the inflation factor. After design is completed, interim analyses of accumulating data can be implemented according to information-based group sequential analysis using the type I error spending function approach, thus obviating the need to conduct interim analyses after equal increments of statistical information and possibly deviating from the planned number of interim analyses and the final analysis, which is unavoidable in complex, large-scale clinical trials in chronic diseases, with nonnormal outcome data.

References

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

Anscombe FJ (1954) Fixed-sample-size analysis of sequential observations. Biometrics 10:89–100

Armitage P (1954) Sequential tests in prophylactic and therapeutic trials. Quart J Med 23:255–274

Armitage P (1975) Sequential medical trials, 2nd edn. Blackwell, Oxford

Armitage P, McPherson CK, Rowe BC (1969) Repeated significance tests on accumulating data. J Roy Stat Soc Ser A 132:235–244

Cox DR (1972) Regression models and life tables (with discussion). J Roy Statist Soc B 34: 187–220

Dillman RO, Seagren SL, Propert KJ et al (1990) A randomized trial of induction chemotherapy plus high-dose radiation versus radiation alone in stage III non-small cell lung cancer. New Engl J Med 323:940–945

Dodge HF, Romig HG (1929) A method of sampling inspection. Bell Syst Tech J 8:613–631

Jennison C, Turnbull BW (1997) Group-sequential analysis incorporating covariate information. J Am Stat Assoc 92:1330–1341

Kim K (1992) Study duration for group sequential trials with censored survival data adjusting for stratification. Stat Med 11:1477–1488

Kim K (1995) SEQPWR and SEQOPR: Computer programs for power calculations and operating characteristics in maximum information trials based on group sequential logrank tests. Comp Meth Progr Biomed 46:143–153

Kim K, DeMets DL (1987) Design and analysis of group sequential tests based on the type I error spending rate function. Biometrika 74:149–154

Kim K, DeMets DL (1992) Sample size determination for group sequential clinical trials with immediate response. Stat Med 11:1391–1399

Kim K, Tsiatis AA (1990) Study duration for clinical trials with survival response and early stopping rule. Biometrics 46:81–92

Kim K, Boucher H, Tsiatis AA (1995) Design and analysis of group sequential logrank tests in maximum duration versus information trials. Biometrics 51:988–1000

Kim K, Tsiatis AA, Mehta CR (2003) Computational issues in information-based group sequential clinical trials. J Jpn Soc Comput Stat 15:153–167

Lan KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. Biometrika 70:659–663

Lan KKG, DeMets DL (1989) Group sequential procedures: Calendar versus information time. Stat Med 8:1191–1198

Lan KKG, Lachin JM (1990) Implementation of group sequential logrank tests in a maximum duration trial. Biometrics 46:759–770

McPherson CK, Armitage P (1971) Repeated significance tests on accumulating data when the null hypothesis is not true. J Roy Stat Soc Ser A 134:15–26

O'Brien PC, Fleming TR (1979) A multiple testing procedure for clinical trials. Biometrics 35:549–556

Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. Biometrika 64:191–199

Propert KJ, Kim K (1992) Group sequential methods in multi-institutional cancer clinical trials: A case study. In: Peace KE (eds) Biopharmaceutical sequential statistical applications. Marcel Dekker, New York, pp 133–153

Reboussin DM, DeMets DL, Kim K, Lan KKG (2000) Computations for group sequential boundaries using the Lan-DeMets spending function method. Contr Clin Trials 21:190–207

Scharfstein DO, Tsiatis AA, Robins JM (1997) Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. J Am Stat Assoc 92:1342–1350

Slud EV, Wei LJ (1982) Two-sample repeated significance tests based on the modified Wilcoxon statistic. J Am Stat Assoc 77:862–868
Wald A (1947) Sequential analysis. Wiley, New York
Whitehead J (1997) The design and analysis of sequential clinical trials, Rev. 2nd edn. Wiley, Chichester

Chapter 4
Sample Size Reestimation for Confirmatory Clinical Trials

Cyrus R. Mehta

1 Introduction

A phase 3 confirmatory clinical trial occurs at the culminating stage of clinical drug development at a biopharmaceutical company. The cost of the entire drug development program from discovery, through animal models, first-in-human testing, phase 2 testing and, eventually, phase 3 testing, can cost over a 100 million dollars. Regulatory approval for a new molecular entity (NME) hinges on the success of at least two adequate and well controlled phase 3 clinical trials. With so much at stake, it is essential that the phase 3 trial be adequately powered. This is often not the case, however. It has been noted by Dr. Janet Woodcock, the Deputy Commissioner at the Center for Drug Evaluation Research, FDA (2008) that one in two clinical trials fail in phase 3 testing. Some of these failures must of course be attributed to unsafe or ineffective compounds. There is, however, concern that an appreciable fraction of these failures might be due to faulty design. For this reason, the FDA (2004) has included adaptive designs in its Critical Path Initiative; a wake-up call to the biomedical community to stem the alarming decline in the number NME's gaining regulatory approval by use of innovative approaches to clinical drug development. An adaptive clinical trial is one in which the future course of the trial can be altered, based on interim data obtained from the on-going trial. Adaptive changes are of various types, including dose selection, population enrichment, alteration of the randomization fractions, and sample size reestimation. In this article, we focus on sample size reestimation as a way of reducing the risk that a study might fail because it is underpowered.

C.R. Mehta (✉)
Cytel Corporation and Harvard School of Public Health, 675 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: mehta@cytel.com

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7_4,

The standard approach for powering a study is to first estimate the underlying treatment effect for the primary endpoint from past studies in similar settings and use this estimate to compute the sample size needed for adequate power. Often, however, there is uncertainty and debate about the magnitude of this estimate, as also about the variability of the patient population, in the current study. This may be due to limited experience with the new compound, small sample sizes in the previous studies, changes in standard of care, lack of comparability between the old and new patient populations, use of a different primary endpoint, and numerous other factors. It is sometimes suggested that rather than attempt to estimate the treatment effect prior to launching the study, it might be more useful to specify the smallest treatment effect that is considered clinically meaningful and power the study to detect this effect. This approach too has its drawbacks. Even assuming that it is possible to identify the smallest clinically meaningful effect, the sample size required to detect that effect might be unacceptably large thereby rendering the trial infeasible. Designing for the smallest clinically meaningful effect has the additional drawback that if the true treatment effect is larger, the trial might be substantially overpowered. One option is to utilize the group sequential design whereby the trial will be terminated early with a smaller sample size if in truth the treatment effect is larger than the smallest clinically meaningful difference. In this article, we present an alternative approach that could be used either in place of, or in conjunction with, the group sequential approach. This is the adaptive approach, wherein an initial sample size is specified on the basis of all available prior information. An interim analysis is then conducted and if the results are reasonably promising, the sample size is increased. In this method, the risk of an underpowered study is reduced since there is an opportunity to review the initial sample size and adjust it based on data from the study itself.

The major portion of this article consists of three case studies. These are examples of confirmatory phase 3 adaptive clinical trials in which we were involved right from the design stage. All three designs were approved by the FDA. For reasons of confidentiality, the names of the trial sponsors and the medical compounds being tested have not been divulged. We introduce the basic notation, statistical methodology, and potential benefit of the adaptive approach in Sect. 2 through a schizophrenia trial with a continuous endpoint. In Sect. 3, we show how the method can be used to extend a very large group sequential design for a cardiology trial with a binomial endpoint. In Sect. 4, we extend the methodology to a lung cancer study in which the primary endpoint is overall survival. Finally, we discuss a number of important statistical and operational issues and summarize our conclusions in Sect. 5.

2 Negative Symptoms Schizophrenia Trial

Consider a two-arm trial to determine if there is an efficacy gain for an experimental drug relative to the clinical standard treatment for negative symptoms schizophrenia. The primary endpoint is the improvement from baseline to week 26 in the Negative

Symptoms Assessment (NSA), a 16-item clinician-rated instrument for measuring the negative symptomatology of schizophrenia. Let, μ_t denote the difference between the mean NSA at baseline and the mean NSA at week 26 for the treatment arm and let μ_c denote the corresponding difference of means for the control arm. Denote the efficacy gain by $\delta = \mu_t - \mu_c$. The trial will be designed to test the null hypothesis H_0: $\delta = 0$ versus the one-sided alternative hypothesis that $\delta > 0$. It is expected, from limited data on related studies, that $\delta \geq 2$ and σ, the between-subject standard deviation, is believed to be about 7.5. In the discussion that follows, we shall focus our attention on adaptive sample size adjustments due to uncertainty surrounding the true value of δ. Even though the statistical methods discussed here are applicable when there is uncertainty about either δ or σ, the adaptive approach requires careful justification primarily when δ is involved. Adaptive sample size adjustments relating to uncertainty about σ are fairly routine and noncontroversial.

We shall consider fixed-sample, group sequential and adaptive design options for this study. There are advantages and disadvantages to each option with no single approach dominating over the others. We are interested, however, in exploring whether the adaptive methodology can add value to the better established fixed sample and group sequential approaches to trial design. We will see that an adaptive design alleviates to some extent the problem of "overruns" encountered by group sequential designs when the primary endpoint is observed after a lengthy follow-up period as is the case here. Additionally, we will see that an adaptive design may, in certain settings, have a more favorable risk versus benefit trade-off.

2.1 Fixed Sample Design

Since it is believed a priori that $\delta \geq 2$, we first create Plan 1, a single-look design with $1-\beta = 0.8$ power to detect $\delta = 2$ using a one-sided level $\alpha = 0.025$ test, given $\sigma = 7.5$. The required sample size can be obtained from the well-known formula

$$N = \sigma^2 \left[\frac{z_\alpha + z_\beta}{\delta}\right]^2, \tag{4.1}$$

where $z_u = \Phi^{-1}(1-u)$ and $\Phi(\cdot)$ is the standard normal cumulative density function. By substituting these design parameters into (4.1), it can be seen that Plan 1 will be fully powered if a total of 441 subjects are enrolled. Efficacy is declared if the standardized test statistic at the time of the final analysis equals or exceeds 1.96.

There is, however, considerable uncertainty about the true value of δ, and to a lesser extent about σ. Nevertheless, it is believed that even if the true value of δ were as low as 1.6 on the NSA scale, that would constitute a clinically meaningful effect. We, therefore, create Plan 2, a one-sided level 0.025 test to detect $\delta = 1.6$ with 0.8 power. Upon substituting the new value of δ into (4.1), it is seen that Plan 2 requires a total sample size of 690 subjects.

Table 4.1 Operating characteristics of Plan 1 and Plan 2

	Plan 1		Plan 2	
δ	Sample size	Power (%)	Sample size	Power (%)
1.6	441	61	690	80
1.7	441	66	690	85
1.8	441	71	690	88
1.9	441	76	690	91
2.0	441	80	690	93

The operating characteristics of Plan 1 and Plan 2 are displayed side by side in Table 4.1 for values of δ between 1.6 and 2.0. Under Plan 1, we would enroll 441 subjects and hope that the study is adequately powered, which it will be if $\delta = 2$ and $\sigma = 7.5$. If, however, $\delta = 1.6$, the power drops from 80 to 61%. There is thus a risk of launching an underpowered study for an effective drug under Plan 1. Under Plan 2, we will enroll 690 subjects, thereby ensuring 80% power at the smallest clinically meaningful value, $\delta = 1.6$, and rising to 93% power at $\delta = 2$.

If resources were plentiful, Plan 2 would clearly be the preferred option. The sponsor must, however, allocate scarce resources over a number of studies and in any case is not in favor of designing an overpowered trial. This leads naturally to considering a design that might be more flexible with respect to sample size than either of the above two single-look fixed sample designs. We will consider two types of flexible designs; group sequential and adaptive.

2.2 Group Sequential Design

Sample size flexibility for late-stage trials is traditionally provided by utilizing a group sequential design. Consider Plan 3, a group sequential design with one interim analysis at which the trial may be terminated early for efficacy. Such a design can be constructed to have 0.025 one sided type-1 error and 80% power to detect $\delta = 1.6$ – the same as Plan 2. If, however, $\delta = 2$, the trial has a high probability of stopping at the interim look itself, rather than going all the way to the end. While this would appear to be an attractive option, it is important to consider not just the saving in study duration but also the saving in the actual number of subjects randomized to the study. Since the efficacy endpoint for this trial will only be observed at week 26, the actual saving in sample size will be affected by the enrollment rate. In the current study, it is anticipated that subjects will enroll at an average rate of 8 per week. The number of subjects enrolled and the number of 26-week completers over time are displayed graphically in Fig. 4.1.

Observe that there is a 26-week horizontal separation between the two parallel lines depicting, respectively, the graph for enrollment and the graph for study completion. This 26-week gap must be taken into consideration when evaluating the savings achieved by utilizing a group sequential design.

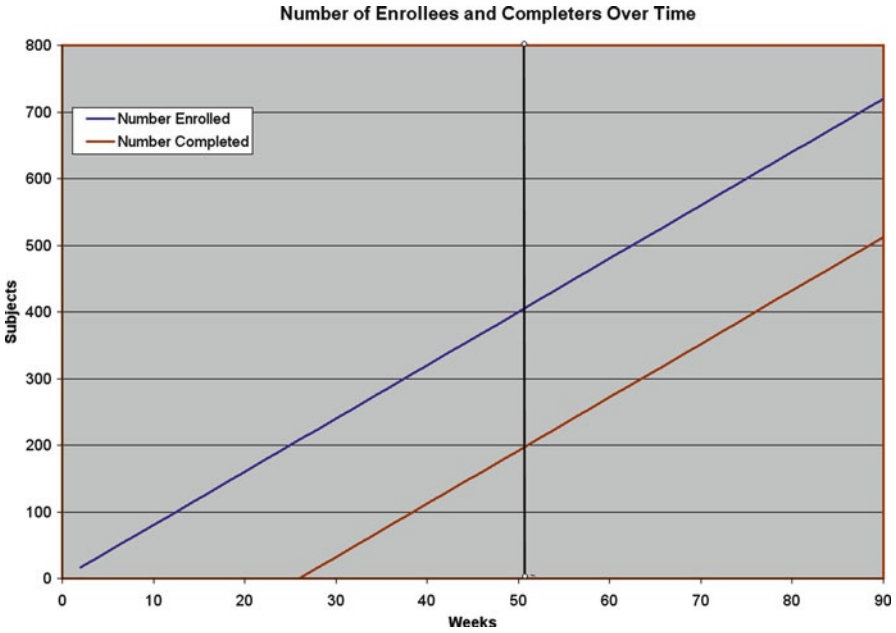

Fig. 4.1 Impact of enrollment rate and length of follow-up on trial completion

The two major design parameters to be specified for a two-look group sequential design are the timing of the interim analysis and the amount of type-1 error to be spent at that analysis. It was felt that data must be available for at least 200 completers before the trial can be terminated for efficacy so that an adequate safety profile may be developed for the study drugs. Therefore, a suitable time point for the interim analysis is week 51, when we will have enrolled 408 subjects with data on 200 completers. Next, we must decide on the amount of type-1 error, α_1, to spend at the interim look. It is generally held that the type-1 error should be spent conservatively in the early stages of a trial so as to ensure that results based on premature termination will be compelling and have the capacity to alter medical practice (see Pocock 2005). A popular choice for this purpose is the proposal of Haybittle (1971) to spend only 0.001 of the total available type-1 error at the interim look. This results in the conservative early stopping boundary $b_1 = \Phi^{-1}(1 - 0.001) = 3.09$. In other words, if the standardized test statistic Z_1 equals or exceeds 3.09 after 26-week data are observed on the first 200 subjects, the trial may be stopped and efficacy declared. The overall type-1 error of the group sequential design is preserved by adjusting the stopping boundary b_2 at the final analysis so as to satisfy the level condition

$$\Pr(Z_1 \geq 3.09) + \Pr(Z_1 < 3.09, Z_2 \geq b_2) = 0.025, \tag{4.2}$$

Table 4.2 Operating characteristics of Plan3 (group sequential) and Plan2 (fixed sample)

	Plan3 (group sequential)					
	Probability of	Expected sample size			Plan2 (fixed sample)	
δ	early stopping (%)	No overruns	With overruns	Power (%)	Sample size	Power (%)
1.6	5.7	665	677	80	690	80
1.7	7.1	658	673	85	690	85
1.8	8.6	651	669	88	690	88
1.9	10.1	643	664	91	690	91
2.0	11.7	635	660	94	690	94

where Z_2 is the standardized test statistic at the final analysis after all enrolled subjects have completed the full 26-week treatment period. The value of b_2 that satisfies (4.2) is $b_2 = 1.97$, slightly larger than the critical value 1.96 for a level-0.025 single look design. This more stringent critical value is sometimes regarded as the penalty one must pay for the privilege of taking an interim look with the possibility of early stopping. It results in a slight inflation of sample size, from 690 to 693, to match the 80% power of the single-look design.

To summarize, in Plan 3 the nominal critical value for early stopping is 3.09 standard deviations and the one sided p-value corresponding to this early stopping boundary 0.001 which, if met, would indeed be compelling enough to justify premature termination. Both Plan 2 and Plan 3 have the same power and almost the same sample size commitment. However, under Plan 2, there is no possibility of early stopping, whereas under Plan 3, it is possible to stop early and thereby save on sample size. The probability of early stopping if $\delta = 1.6$ is about 6%, and rises to about 12% if $\delta = 2$. Thus, the benefit of the group sequential design in this setting appears to be small and moreover is accompanied by the problem of "overruns" which we discuss next.

2.2.1 The Problem of Overruns

Care must be taken when estimating the actual sample size savings of a group sequential design. Even if the early stopping boundary is crossed at week 51 on the basis of the data from the 200 completers, we must still take into account the additional 208 randomized subjects who enrolled between week 26 and 51 for whom the week 26 endpoint will not yet have been attained. These additional 208 subjects are referred to as the "overruns." When the overruns are accounted for, the saving in sample size due to early stopping is only $693 - 408 = 285$ subjects, rather than $693 - 200 = 493$ subjects. The power and expected sample size values of the group sequential Plan 3 for different choices of δ are displayed in Table 4.2. The table shows the impact of overruns on the expected sample size. For comparison, we have also included corresponding power and sample size values for the fixed sample Plan 2 in Table 4.2.

It is seen from Table 4.2 that Plan 3 offers a modest benefit relative to Plan 2. After accounting for the overruns, the expected sample sizes under Plan 3 range between 660 and 677 for corresponding values of δ between 2 and 1.6, as compared to a fixed sample size of 690 under Plan2. In terms of power, Plan 2 and Plan 3 are practically identical. For the current trial, a group sequential design with conservative error spending offers no substantial advantage over a conventional single look design with a fixed sample size. One is still faced with the dilemma of committing excessive sample size resources up front to ensure adequate power at $\delta = 1.6$, with limited prospects of saving on sample size in the event that $\delta = 2$.

Although in general group sequential designs do offer savings in expected sample size, their actual benefit may be diminished if a study enrolls subjects very rapidly but the primary endpoint can only be observed after a lengthy follow-up. In the current example, we assumed that subjects are enrolled at the rate of 8 per week and the endpoint is observed after 26 weeks of follow-up for each subject. This resulted in 208 additional subjects being on-study who were not yet followed for 26 weeks at the time of the interim analysis. The efficiency loss due to an overrun of this magnitude was difficult to overcome. If instead the enrollment rate were to be halved to four subjects per week, and the endpoint were to be observed only after 12 weeks instead of 26 weeks, there would only be an overrun of 48 subjects, and the resulting operating characteristics of the two group sequential designs would be more favorable relative to the corresponding fixed sample design. The accrual rate and the duration of follow-up are thus two extremely important design parameters for a group sequential trial.

From an operational point of view, overruns pose an additional problem. If a trial is terminated early, its results are typically unblinded to the investigators. In this case, there is the risk that the investigators might no longer feel bound to treat the overrun patients strictly as per protocol. With a large number of overruns, the integrity of the study might be compromised because the dataset used for the final analysis, unlike the dataset used for the early stopping decision, must include all available data on all randomized subjects.

We next consider adopting an adaptive design for this study. This is a radically different approach to trial design in which the difficulties encountered by group sequential designs – rapid accrual, delayed endpoint, and large up-front commitment of patient resources – can to some extent be mitigated.

2.3 Adaptive Design

To motivate the adaptive design, let us recall that although the actual value of δ is unknown, the investigators are hopeful that $\delta \geq 2$. For this reason, Plan 1 was constructed to have 80% power to detect $\delta = 2$. Plan 2, on the other hand, was constructed to have 80% power to detect $\delta = 1.6$, the smallest clinically meaningful treatment effect. If there were no resource constraints, one would of course prefer to design the study for 80% power at $\delta = 1.6$ since that would imply

even more power at $\delta = 2$. However, as we saw in Table 4.1, this conservative strategy carries as its price a substantially larger up-front sample size commitment which is, moreover, unnecessary if in truth $\delta = 2$. Plan 3 was, therefore, constructed as a group sequential alternative to Plan 2. Plan 3 also has 80% power to detect $\delta = 1.6$ but there is a possibility of early stopping. We have seen, however, that due to the overruns problem and the need to impose rather conservative criteria for early stopping, the expected sample size saving realized by Plan 3 is small while the up-front sample size commitment is large.

The above difficulties lead us to consider whether Plan 1, though powered to detect $\delta = 2$, might be improved so as to provide some insurance against substantial power loss in the event that $\delta = 1.6$. The adaptive approach is suited to this purpose. In this approach, we start out with a sample size of 441 subjects as in Plan 1, but take an interim look after data are available on 200 completers. The purpose of the interim look is not to stop the trial early but rather to examine the interim data and continue enrolling past the planned 441 subjects if the interim results are promising enough to warrant the additional investment of sample size. This strategy has the advantage that the sample size is finalized only after a thorough examination of data from the actual study rather than through making a large up-front sample size commitment before any data are available. Furthermore, if the sample size may only be increased but never decreased from the originally planned 441 subjects, there is no loss of efficiency due to overruns. An important paper by Proschan and Hunsberger (1995) has shown, however, that such a data dependent sample size increase will inflate the type-1 error unless appropriate adjustments are made at the time of the final analysis. This is discussed next.

2.3.1 Preserving the Type-1 Error

The technical problem of avoiding inflating the type-1 error despite increasing the sample size in a data-dependent manner has been approached from two different points of view. One approach is based on combining the independent data from the two stages using a prespecified combination function. The first papers to develop this approach were by Bauer and Köhne (1994), Cui et al. (1999), and Lehmacher and Wassmer (1999). Bauer and Köhne (1994) proposed to combine the p-values from the two stages multiplicatively, in a manner analogous to meta-analysis (see, e.g., Hedges and Olkin 1985). Thus, if $p^{(1)}$ and $p^{(2)}$ are valid p-values computed independently from the two stages then, under the null hypothesis of no treatment effect, they are uniformly distributed and their product is distributed as χ^2 with 4 degrees of freedom. This property is preserved even if the sample size at the second stage is altered on the basis of the data observed at the first stage. Cui et al. (1999) proposed to combine the independent standardized (Wald) statistics from the two stages rather than the p-values. They used prespecified weights and an additive combination function. Specifically, suppose that the initial plan is to enroll $n^{(1)}$ subjects at stage 1 and to enroll an independent cohort of $n^{(2)}$ subjects at stage 2. Suppose that on the basis of the interim results at stage 1, the sample size of the

second cohort is increased from $n^{(2)}$ to $n^{(2*)}$. Let $Z^{(1)}, Z^{(2)}$, and $Z^{(2*)}$ denote the Wald statistics for the first stage data, second stage data, and second stage data with altered sample size, respectively. Cui et al. (1999) have shown that the statistic

$$\tilde{Z}_2 = \sqrt{\frac{n^{(1)}}{n^{(1)}+n^{(2)}}}Z^{(1)} + \sqrt{\frac{n^{(2)}}{n^{(1)}+n^{(2)}}}Z^{(2*)} \tag{4.3}$$

is $N(0,1)$ under the null hypothesis. This result is exactly true for normally distributed data with known variance, and asymptotically true for other settings including the binomial approximation to the normal. The significance of (4.3) is that one can perform the final analysis using the criterion $\tilde{Z}_2 \geq z_\alpha$ to reject the null hypothesis and, notwithstanding the data dependent increase in sample size at stage 1, the type-1 error will not be inflated because

$$\Pr(\tilde{Z}_2 \geq z_\alpha) = \alpha \ .$$

Lehmacher and Wassmer (1999) proposed to combine the p-values from the two stage by the inverse normal combination function

$$\tilde{p} = 1 - \Phi\left[\sqrt{\frac{n^{(1)}}{n^{(1)}+n^{(2)}}}\Phi^{-1}\left(1-p^{(1)}\right) + \sqrt{\frac{n^{(2)}}{n^{(1)}+n^{(2)}}}\Phi^{-1}\left(1-p^{(2*)}\right)\right] \tag{4.4}$$

and to reject the null hypothesis if $\tilde{p} \leq \alpha$. Although they developed their results independently, it is easy to see that the Lehmacher and Wassmer (1999) and the Cui et al. (1999) methods are completely equivalent.

Observe from (4.3) and (4.4) that the data from the two stages are combined with weights that depend on their prespecified sample sizes $(n^{(1)}, n^{(2)})$ rather than their actual sample sizes $(n^{(1)}, n^{(2*)})$. When $n^{(2*)} > n^{(2)}$, this implies that the Wald statistic (or associated p-value statistic) from the second cohort is "down-weighted" relative to the corresponding statistic from the first cohort. This has been a source of controversy amongst statisticians both because of concerns that the resulting test might lose some efficiency (Tsiatis and Mehta 2003; Jennison and Turnbull 2003) and because it seems illogical that the data generated by subjects in the second cohort should count less than that of subjects in the first cohort (Chen et al. 2004). We shall return to these issues in Sect. 5 where we argue that their impact on actual clinical trials is rather limited.

An alternative approach to preserving the type-1 error is based on calculating the conditional error rate. This approach underlies the papers by Proschan and Hunsberger (1995) and Müller and Schäfer (2001). The basic principle is that the conditional type-1 error of the trial with the sample size adaptation must remain the same as the conditional type-1 error of the original nonadaptive trial. Let, Z_2 denote the Wald statistic for the cumulative data from both stages if there is no sample size adaptation and Z_2^* denote the corresponding Wald statistic if the sample size of the second cohort is increased from $n^{(2)}$ to $n^{(2*)}$ after observing $Z_1 = z_1$ at the

first stage. The nonadaptive trial will reject the null hypothesis at level-α if $Z_2 \geq z_\alpha$. In order for the adaptive trial to also operate at the same level, its critical value must be adjusted from z_α to $c(z_1)$ such that, under the null hypothesis $\delta = 0$,

$$\Pr(Z_2^* \geq c(z_1)|z_1) = \Pr(Z_2 \geq z_\alpha|z_1)\ . \tag{4.5}$$

The above principle of preserving the conditional error rates can be applied to more general adaptations than sample size increase. However, for the case of sample size increase, it can be shown that the method of preserving error rates and the method of combining the data from the two stages by prespecified weights are completely equivalent (see, e.g., Gao et al. 2008).

2.3.2 Selecting the Criteria for an Adaptive Sample Size Increase

The operating characteristics of an adaptive design depend in a complicated way on the criteria for increasing the sample size after observing the interim data. These criteria may combine objective information such as the current estimate of δ or the current conditional power with assessments of safety and with information available from other clinical trials that was not available at the start of the study. The adaptive approach provides complete flexibility to modify the sample size without having to prespecify a precise mathematical formula for computing the new sample size based on the interim data. However, for confirmatory trials, it is a regulatory requirement that the precise sample-size reestimation rule be prespecified as far as possible, subject of course to being overruled if there are safety concerns or other matters requiring clinical judgement, at the time of the interim analysis. It is thus instructive to investigate power and expected sample size by simulating the trial under different values of δ and applying precise prespecified rules for increasing the sample size on the basis of the observed interim results.

To this end, we create Plan 4, a design with 80% power to detect $\delta = 2$ with a one-sided level-0.025 test, based on a planned enrollment of 441 subjects. Plan 4 specifies, in addition, that there will be one interim analysis after 26 weeks of follow-up data are available on the first 200 subjects enrolled. The purpose of the interim analysis is not to stop the trial early but rather to examine the interim data and decide whether a sample size increase is warranted. If no action were taken at the interim look, Plan 4 would be identical to Plan 1. The timing of the interim look reflects a preference for performing the interim analysis as late as possible but nevertheless while the trial is still enrolling subjects since, once the enrollment sites have closed down, it will be difficult to start them up again. Under the assumption that subjects enroll at the rate of 8 per week, we will have enrolled 408 subjects by week 51; 200 of them will have completed the required 26 weeks of follow-up for the primary endpoint, and an additional 208 subjects will comprise the overruns. Only the data from the 200 completers will be used in making the decision to increase the sample size. After this decision is taken, enrollment will continue until the desired sample size is attained. The primary efficacy analysis will be based on

the full 26 weeks of follow-up data from all enrolled subjects so that there can be no ambiguity regarding the statistical outcome. This may be contrasted with the group sequential setting where the 208 overruns played no direct role in the primary efficacy analysis at the time that an early stopping decision was implemented. Thus, one may face an ambiguous situation should the data produced by these overruns after they become completers lead to a reversal of the conclusion reached at the time of early stopping.

It remains to specify a precise formula for increasing the sample size. While there are an infinite number of ways to construct such a formula, it must address the following three questions:

- For what range of interim outcomes should a sample size increase be contemplated?
- How should the magnitude of the new sample size be calculated?
- What should be the upper limit to the sample size increase?

The answers to these questions might be driven by both clinical and business concerns, and will depend on the importance the investigators place on avoiding a false negative outcome for the current trial.

Range of Interim Outcomes for a Sample Size Increase

It is convenient to partition the sample space of possible interim outcomes into three zones; *unfavorable, promising, and favorable*. An adaptive strategy is built on the premise that if the interim outcome lies in either the unfavorable or favorable zones, it is unnecessary to alter the sample size. In one case, it would be risky to invest further in what appears to be a failed trial, while in the other case the trial appears slated to succeed anyway, without an additional sample size investment. Thus, an adaptive sample size increase is only intended to help studies whose interim results fall in a promising zone, between these two extremes. How might these three zones be identified? One could use the interim estimate $\hat{\delta}$ or its standardized version $z = \hat{\delta}/\mathrm{se}(\hat{\delta})$ to partition the sample space into the three zones. Alternatively, one could rely on the conditional power or *probability of obtaining a positive outcome at the end of the trial, given the data already observed*. The conditional power approach is favored by most practitioners because it has a meaningful interpretation that is independent of the type of endpoint being measured, and incorporates both the current estimate of treatment effect as well as its standard error. Conditional power is a function of both δ, the underlying treatment effect, and z_1, the interim value of the Wald statistic. Since δ is unknown, it is usual and convenient to replace it with the estimate $\hat{\delta}_1$ obtained at the interim analysis. Conditional power is then evaluated by the formula

$$\mathrm{CP}_{\hat{\delta}_1}(z_1) = 1 - \Phi\left(\frac{z_\alpha\sqrt{n^{(1)}+n^{(2)}} - z_1\sqrt{n^{(1)}}}{\sqrt{n^{(2)}}} - \frac{z_1\sqrt{n^{(2)}}}{\sqrt{n^{(1)}}}\right). \tag{4.6}$$

For the present trial, we prespecify that a sample size increase will only be contemplated if the conditional power at the interim look lies between 30 and 80%. That is, the unfavorable zone is characterized by conditional power values at most equal to 30%, the promising zone by conditional power values between 30 and 80% and the favorable zone by conditional power values at least equal to 80%.

Computing the Required Sample Size Increase

Just as at the design stage of a trial, the sample size is determined by the desired power (80%, say) to detect an anticipated value of δ, so also at the time of the interim analysis the new sample size may be determined by the desired conditional power (also 80%, say) to detect an anticipated value of δ. Now, however, data from the actual trial are available, and may be used to update the anticipated value of δ at which to power the trial. One could, if desired, incorporate prior beliefs, external information and current data into a value of δ at which to power the study. For simplicity, however, we shall use the estimate of δ obtained at the interim analysis to recompute the sample size so as to achieve 80% conditional power. The new sample size is obtained by searching for the value $n^{(2*)}$ that will make the right-hand side of (4.6) equal to 0.8. It is possible that this calculation could result in a reduction in the total sample size. This is permitted by the statistical methodology of adaptive designs. For the current example, however, we do not wish to decrease the sample size. Therefore, if the recomputed sample size constitutes a decrease, the original sample size of 441 subjects will be used.

Specifying an Upper Limit to the Sample Size Increase

Since resources are limited, there must be an upper limit to the sample size increase, no matter what sample size is required to attain 80% conditional power. This upper limit is usually restricted to between 50 and 100% of the original sample size and is prespecified at the start of the trial. Larger sample size increases are undesirable since they could yield statistically significant outcomes that are clinically nonsignificant. For the current trial, we prespecify an upper limit of 882 subjects. That is, we are prepared to double our investment in the trial, but only if the interim estimate of conditional power falls in the favorable zone.

Finally, the design specifications of the adaptive Plan 4 are as follows:

1. The initial sample size is 441 subjects, and has 80% power to detect $\delta = 2$ with a one-sided level-0.025 test.
2. An interim analysis is performed after data are available on 200 completers with 26 weeks of follow-up data.
3. At the interim analysis, the conditional power is computed using the estimated value $\hat{\delta}$ as though it were the true value of δ. If the conditional power lies between 30 and 80%, the interim outcome is deemed to be promising.

Table 4.3 Operating characteristics of Plan 1 (fixed sample) and Plan 4 (adaptive)

	Plan 1 (fixed sample)		Plan 4 (adaptive)	
Value of δ	Power (%)	Expected sample size	Power (%)	Expected sample size
1.6	61	441	67	509
1.7	66	441	72	507
1.8	71	441	76	506
1.9	76	441	81	502
2.0	80	441	84	499

All Plan 4 results are based on 100,000 simulated trials

4. If the interim outcome is promising, the sample size is recomputed so as to achieve 80% conditional power at the estimated value, $\hat{\delta}$. The original sample size is then updated to the recomputed sample size, subject to the constraint in item 5 shown below.
5. If the recomputed sample size is less than 441, the original sample size of 441 subjects is used. If the recomputed sample size exceeds 882, the sample size is curtailed to 882 subjects.

2.3.3 Operating Characteristics of Adaptive Design

Due to the complex adaptive scheme for recomputing sample size, the operating characteristics of Plan 4 can best be evaluated by simulation. Table 4.3 displays power and expected sample size values for selected values of δ between 1.6 and 2.0, based on 100,000 simulations of Plan 4. For comparative purposes, corresponding power and sample size values for Plan 1 are also displayed.

The power of the adaptive Plan 4 has increased by 6% at $\delta = 1.6$ and by 4% at $\delta = 2$ compared to Plan 1. These power gains were obtained at the cost of corresponding average sample size increases of 67 subjects at $\delta = 1.6$ and 57 subjects at $\delta = 2$. The gains in power appear to be fairly modest, especially as they are offset by corresponding sample size increases. However, Plan 4 offers a significant benefit in terms of risk reduction, not reflected in Table 4.3. To see this, it is important to note that the sample size under Plan 4 is only increased when the interim results are promising; that is, when the conditional power at the interim analysis is greater than 30% but less than 80%. This is the very situation in which it is advantageous to increase the sample size and thereby avoid an underpowered trial. When the interim results are unfavorable (conditional power $\leq 30\%$) or favorable (conditional power $\geq 80\%$), a sample size increase in not warranted and hence the sample size is unchanged at 441 subjects for both Plan 1 and Plan 4. But when the interim results are promising (conditional power between 30 and 80%) the sample size is increased under Plan 4 in an attempt to boost the conditional power back to 80%. It is this feature of the adaptive design that makes it more attractive than the simpler fixed sample design.

Table 4.4 displays the probability of falling into the unfavorable, promising, and favorable zones at the interim look, along with the power and expected sample

Table 4.4 Operating characteristics of Plan 1 and Plan 4 conditional on interim outcome

δ	Interim outcome	Probability of interim outcome (%)	Power conditional on interim outcome		Expected sample size	
			Plan 1 (%)	Plan 4 (%)	Plan 1	Plan 4
1.6	Unfavorable	33	27	27	441	441
	Promising	26	61	83	441	696
	Favorable	41	87	87	441	441
1.7	Unfavorable	29	32	32	441	441
	Promising	26	65	86	441	694
	Favorable	45	90	90	441	441
1.8	Unfavorable	26	36	36	441	441
	Promising	26	69	89	441	690
	Favorable	48	91	91	441	441
1.9	Unfavorable	23	41	41	441	441
	Promising	25	72	91	441	689
	Favorable	52	93	93	441	441
2.0	Unfavorable	20	44	44	441	441
	Promising	24	76	93	441	685
	Favorable	56	95	95	441	441

All results are based on 100,000 simulated trials

size, conditional on falling into each zone, under both Plan 1 and Plan 4. The table highlights the key advantage of the adaptive Plan 4 compared to the fixed sample Plan 1; that is, the ability to invest in the trial in stages, with the second stage of the investment being required only if promising results are obtained at the first stage. This feature of Plan 4 makes it far more attractive as an investment strategy than Plan 1 which has no provision for increasing the sample size if a promising interim outcome is obtained. Suppose, for example, that $\delta = 1.6$, the smallest clinically meaningful treatment effect. The trial sponsor only commits the resources needed for 441 subjects at the start of the trial, at which point the chance of success is 61%, as shown in Table 4.3. The additional sample size commitment is forthcoming only if promising results are obtained at the interim analysis, and in that case the sponsor's risk is substantially reduced because the chance of success jumps to 83%, as shown in Table 4.4. Similar results are observed for the other values of δ.

The probabilities of entering the unfavorable, promising, and favorable zones at the interim analysis, displayed in Table 4.4, are instructive. Consider again the case $\delta = 1.6$. At this value of δ, there is a 26% chance of landing in the promising zone and thereby obtaining a substantial power boost under Plan 4 as compared to Plan 1. That is, 26% of the time the adaptive strategy can rescue a trial that is underpowered at the interim look. The chance of entering the favorable zone is 41%. That is, 41% of the time the sponsor will be lucky and have a well powered trial at the interim look without the need to increase the sample size. The remaining 33% of the time, the sponsor will be unlucky and will enter the unfavorable zone from which there is no sample size increase, and the chance of success is only 27%. These odds improve with larger values of δ.

3 Acute Coronary Syndromes

We designed a two-arm, placebo controlled randomized clinical trial for subjects with acute cardiovascular disease undergoing percutaneous coronary intervention (Mehta et al. 2007). The primary endpoint is a composite of death, myocardial infarction, or ischemia-driven revascularization during the first 48 h after randomization. We assume on the basis of prior knowledge that the event rate for the placebo arm is 8.7%. The investigational drug is expected to reduce the event rate by at least 20%. The investigators are planning to randomize a total of 8,000 subjects in equal proportions to the two arms of the study. It is easy to show that a conventional fixed sample design enrolling a total of 8,000 subjects will have 83% power to detect a 20% risk reduction with a one-sided level-0.025 test of significance. The actual risk reduction is expected to be larger, but could also be as low as 15%, a treatment effect that would still be of clinical interest given the severity and importance of the outcomes. In addition, there is some uncertainty about the magnitude of the placebo event rate. For these reasons, the investigators wish to build into the trial design some flexibility for adjusting the sample size. Two options under consideration are, a group sequential design with the possibility of early stopping in case the risk reduction is large, and an adaptive design with the possibility of increasing the sample size in case the risk reduction is small. In the remainder of this section, we shall discuss these two options and show how they may be combined into a single design that captures the benefits of both.

3.1 Group Sequential Design

We first transform the fixed sample design into an 8,000 person group sequential design with two interim looks, one after 4,000 subjects are enrolled (50% of total information) and the second after 5,600 subjects are enrolled (70% of total information). Early stopping efficacy boundaries are derived from the Lan and DeMets (1983) O'Brien-Fleming type error spending function. Let us denote this group sequential design as GSD1. The operating characteristics of GSD1 are displayed in Table 4.5. The first column of Table 4.5 is a list of potential risk reductions, defined as $100 \times (1 - \rho)\%$ where $\rho = \pi_t / \pi_c$, π_t is the event rate for the treatment arm, and π_c is the event rate for the control arm. The remaining columns display early stopping probabilities, power and expected sample size. Since the endpoint is observed within 48 h, the problem of overruns that we encountered in the schizophrenia trial is negligible and may be ignored.

Table 4.5 shows that GSD1 is well powered, with large savings of expected sample size for risk reductions of 20% or more. It is thus a satisfactory design if, as is initially believed, the magnitude of the risk reduction is in the range 20–25%. This design does not, however, offer as good protection against a false negative conclusion for smaller risk reductions. In particular, even though 15% is

Table 4.5 Operating characteristics of GSD1, a three-look 8,000-person group sequential design

	Probability of crossing efficacy boundary				
Risk reduction $100 \times (1-\rho)$ (%)	At look 1 ($N = 4{,}000$)	At look 2 ($N = 5{,}600$)	At final look ($N = 8{,}000$)	Overall power (%)	Expected sample size
15	0.074	0.183	0.309	57	7,264
17	0.109	0.235	0.335	68	7,002
20	0.181	0.310	0.330	82	6,535
23	0.279	0.362	0.275	92	6,017
25	0.357	0.376	0.222	96	5,671

Table 4.6 Operating characteristics of GSD2, a three-look 13,853-person Grp sequential design

	Probability of crossing efficacy boundary				
Risk reduction $100 \times (1-\rho)$ (%)	At look 1 ($N = 6{,}926$)	At look 2 ($N = 9{,}697$)	At final look ($N = 13{,}853$)	Overall power (%)	Expected sample size
15	0.167	0.298	0.335	80	11,456
17	0.246	0.349	0.296	89	10,699
20	0.395	0.375	0.196	97	9,558
23	0.565	0.329	0.099	99.3	8,574
25	0.675	0.269	0.054	99.8	8,061

still a clinically meaningful risk reduction, GSD1 offers only 57% power to detect this treatment effect. One possibility then is to increase the up-front sample size commitment of the group sequential design so that it has 80% power if the risk reduction is 15%. This leads to GSD2, a three-look group sequential design with a maximum sample size commitment of 13,853 subjects, one interim look after 6,926 subjects (50% of total information) and a second interim look after 9,697 subjects (70% of total information). GSD2 has 80% power to detect a risk reduction of 15% with a one-sided level-0.025 test.

Table 4.6 displays operating characteristics of GSD2 for risk reductions between 15 and 25%. Notice that by attempting to provide adequate power at 15% risk reduction, the low end of clinically meaningful treatment effects, we have significantly over-powered the trial for values of risk reduction in the expected range of risk reductions, 20–25%. If, as expected, the risk reduction exceeds 20%, the large up-front sample size commitment of 13,853 subjects under GSD2 is unnecessary. GSD1 with an up-front commitment of only 8,000 subjects will provide sufficient power in this setting. From this point of view, GSD2 is not a very satisfactory design. It commits the investigators to a very large and expensive trial to provide adequate power in the pessimistic range of risk reductions, without any evidence that the true risk reduction does indeed lie in the pessimistic range. Evidently, a single group sequential design cannot provide adequate power for the "worst-case" scenario, and at the same time avoid overpowering the more optimistic range of scenarios. This leads us to consider building an adaptive sample size reestimation option into the group sequential design GSD1, such that the adaptive component will provide the necessary insurance

for the worst-case scenario, and thereby free the group sequential component to provide adequate power for the expected scenario, without a large and unnecessary up-front sample size commitment.

3.2 Adaptive Group Sequential Design

We convert the three-look group sequential design GSD1 into an adaptive group sequential design by inserting into it the option to increase the sample size at look 2, when 5,600 subjects have been enrolled. Denote the modified design by A-GSD1. The rules governing the sample size increase for A-GSD1 are similar to the rules specified in Sect. 2 for the schizophrenia trial, but tailored to the needs of the current trial. The idea is to identify unfavorable, promising and favorable zones for the interim results at look 2, based on the attained conditional power. Sample size should only be increased if the interim results fall in the promising zone. Subject to an upper limit, the sample size should be increased by just the right amount to boost the current conditional power to some desired level (say 80%). The following are the design specifications for A-GSD1:

1. The starting design is GSD1 with a sample size of 8,000 subjects, one interim look after enrolling 4,000 subjects and a second interim look after enrolling 5,600 subjects. The efficacy stopping boundaries at these two interim looks are derived from the Lan and DeMets (1983) error spending function of the O'Brien-Fleming type.
2. At the second interim analysis, with data available on 5,600 subjects, the conditional power is computed using the estimated value $\hat{\rho}$ as though it were the true relative risk ρ. If the conditional power is no greater than 30% the outcome is deemed to be unfavorable. If the conditional power is between 30 and 80%, the outcome is deemed to be promising. If the conditional power is at least 80%, the outcome is deemed to be favorable.
3. If the interim outcome is promising, the sample size is recomputed so as to achieve 80% conditional power at the estimated value $\hat{\rho}$. The original sample size is then updated to the recomputed sample size, subject to the constraint in item 4 shown below.
4. If the recomputed sample size is less than 8,000, the original sample size of 8,000 subjects is used. If the recomputed sample size exceeds 16,000, the sample size is curtailed at 16,000 subjects.

Some features of this adaptive strategy are worth pointing out. First, the sample size is recomputed on the basis of data from 5,600 subjects from the trial itself. Therefore, the estimate of ρ available at the interim analysis is substantially more reliable than the estimate that was used at the start of the trial to compute an initial sample size of 8,000 subjects. The latter estimate is typically derived from smaller pilot studies or from other phase 3 studies in which the patient population might not be exactly the same as that of the current trial. Second, a sample size increase is only

Table 4.7 Operating characteristics of GSD1 (group sequential) and A-GSD1 (adaptive group sequential) designs

Risk reduction	GSD1 (group sequential)		A-GSD1 (adaptive group sequential)	
$100 \times (1 - \rho)$ (%)	Power (%)	Expected sample size	Power (%)	Expected sample size
15	57	7,264	63	8,357
17	68	7,002	74	8,010
20	82	6,535	86	7,324
23	92	6,017	94	6,496
25	96	5,671	97	5,989

All results for A-GSD1 are based on 100,000 simulated trials

requested if the interim results are promising, in which case the trial sponsor should be willing to invest the additional resources needed to power the trial adequately. In contrast, GSD2 increases the sample size substantially at the very beginning of the trial, before any data are available to determine if the large sample size is justified.

3.3 Operating Characteristics of Adaptive Group Sequential Design

Table 4.7 displays the power and expected sample size of the adaptive group sequential design A-GSD1. For comparative purposes, corresponding power and sample size values of GSD1 are also provided. If there is a 15% risk reduction, A-GSD1 has 6% more power than GSD1 but utilizes an additional 1,093 subjects on average. It is seen that as the risk reduction parameter increases, the power advantage and additional sample size requirement of A-GSD1 are reduced relative to GSD1.

The power and sample size entries in Table 4.7 were computed unconditionally, and for that reason do not reveal the real benefit that design A-GSD1 offers compared to design GSD1. As discussed previously in the schizophrenia example, the real benefit of an adaptive design is the opportunity it provides to invest in the trial in stages with the second stage investment forthcoming only if promising results are obtained at the first stage. To explain this better, it is necessary to display power and expected sample size results conditional on the zone (unfavorable, promising or favorable) into which the results of the trial fall at the second interim analysis. Accordingly, Table 4.8 displays the operating characteristics of both GSD1 and A-GSD1 conditional on the zone into which the conditional power falls at the second interim analysis. The table reveals substantial gains in power for A-GSD1 compared to GSD1 at all values of risk reduction if the second interim outcome falls in the promising zone, thereby leading to an increase in the sample size. Outside this zone, the two designs have the same operating characteristics since the sample size does not change. If the second interim outcome falls in the unfavorable zone, the trial appears to be headed for failure and an additional sample size investment would be

Table 4.8 Operating characteristics of GSD1 (group sequential) and A-GSD1 (adaptive group sequential) designs conditional on second interim outcome

Risk reduction	Second interim	Probability of interim	Power conditional on second interim outcome		Expected sample size	
$100 \times (1-\rho)$ (%)	outcome	outcome (%)	GSD1 (%)	A-GSD1 (%)	GSD1	A-GSD1
15	Unfavorable	36	15	15	8,000	8,000
	Promising	24	57	81	8,000	12,068
	Favorable	40	94	94	6,152	6,160
17	Unfavorable	27	19	20	8,000	8,000
	Promising	24	64	87	8,000	11,938
	Favorable	50	96	96	5,992	5,989
20	Unfavorable	16	29	29	8,000	8,000
	Promising	20	73	93	8,000	11,758
	Favorable	64	98	98	5,721	5,721
23	Unfavorable	9	40	40	8,000	8,000
	Promising	14	81	96	8,000	11,553
	Favorable	77	99	99	5,440	5,435
25	Unfavorable	5	48	48	8,000	8,000
	Promising	11	85	98	8,000	11,465
	Favorable	84	99.6	99.5	5,250	5,238

All results are based on 100,000 simulated trials

risky. If the second interim outcome falls in the favorable zone, the trial is headed for success without the need to increase the sample size. Thus the adaptive design provides the opportunity to increase the sample size only when the results of the second interim analysis fall in the promising zone. This is precisely when the trial can most benefit from a sample size increase.

3.4 Adding a Futility Boundary

One concern with design A-GSD1 is that it lacks a futility boundary. There is thus the risk of proceeding to the end, possibly with a sample size increase, when the magnitude of the risk reduction is small and unlikely to result in a successful trial. In particular, suppose that the null hypothesis is true. In that case, we can show that the power (i.e., the type-1 error) is 2.5% and the expected sample size under A-GSD1 is 8,256 subjects. It might thus be desirable to include some type of futility stopping rule for the trial. In this trial, the investigators proposed the following futility stopping rules at the two interim analysis time points:

1. Stop for futility at the first interim analysis ($N = 4{,}000$) if the estimated event rate for the experimental arm is at least 1% higher than the estimated event rate for the control arm.
2. Stop for futility at the second interim analysis ($N = 5{,}600$) if the conditional power, based on the estimated risk ratio $\hat{\rho}$, is no greater than 20%.

Table 4.9 Operating characteristics of the A-GSD1 design with and without a futility boundary

Risk reduction	A-GSD1 with no futility boundary		A-GSD1 with futility boundary	
$100 \times (1-\rho)$ (%)	Power (%)	Expected sample size	Power (%)	Expected sample size
0	2.5	8,293	2.5	5,411
15	63	8,357	59	7,438
20	86	7,324	83	6,839
25	96	5,989	94	5,897

All results are based on 100,000 simulated trials

Table 4.10 Operating characteristics of A-GSD1 design with and without a futility boundary, conditional on the second interim outcome

Risk reduction	Second interim	Probability of interim	Power conditional on second interim outcome		Expected sample size	
$100 \times (1-\rho)$ (%)	outcome	outcome (%)	No fut (%)	With fut (%)	No fut	With fut
0	Unfavorable	92	0.5	0.1	8,000	4,850
	Promising	6	15	15	12,926	12,788
	Favorable	2	65	61	6,911	7,053
15	Unfavorable	36	15	5	8,000	5,711
	Promising	24	81	80	12,068	11,892
	Favorable	40	94	93	6,160	6,207
20	Unfavorable	16	29	10	8,000	5,910
	Promising	20	93	92	11,758	11,637
	Favorable	64	98	98	5,721	5,765
25	Unfavorable	5	48	17	8,000	6,100
	Promising	11	98	97	11,465	11,345
	Favorable	84	99.5	99.5	5,238	5,268

All results are based on 100,000 simulated trials

The impact of the futility boundary on the unconditional operating characteristics of the A-GSD1 design are displayed in Table 4.9. The inclusion of the futility boundary has resulted in a dramatic saving of nearly 3,000 subjects, on average, at the null hypothesis of no risk reduction. Furthermore, notwithstanding a small power loss of 2–3%, the trial continues to have well over 80% power for risk reductions of 20% or more. The trial suffers a power loss of 4% if the magnitude of the risk reduction is 15%, the low end of the range of clinical interest. In this situation, however, the unconditional power is inadequate (only 63%) even without a futility boundary. To fully appreciate the impact of the futility boundary on power and expected sample size, it is necessary to study the operating characteristics of the trial conditional on the results of the second interim analysis. These results are displayed in Table 4.10. It is seen that the presence of the futility boundary does not cause any loss of power for trials that enter the promising or favorable zones at the second interim analysis. Additionally, the presence of the futility boundary causes the average sample size to be reduced substantially in the unfavorable zone while remaining the same in the other two zones. In effect, the futility boundary terminates a proportion of trials that enter the unfavorable zone thereby preventing them from proceeding to conclusion. It has no impact on trials that enter the promising or favorable zones.

4 Nonsmall Cell Lung Cancer

A two arm double-blind multi-center randomized clinical trial was recently initiated for subjects with advanced metastatic nonsmall cell lung cancer with the goal of comparing the industry standard control therapy to a new therapy. The primary endpoint was overall survival (OS). The study was powered to detect a 25% improvement in median survival, from 7.4 months on the control arm to 9.25 months on the experimental arm, which corresponds to a hazard ratio of 0.8. A group sequential design was adopted with an efficacy boundary derived from the Lan and DeMets (1983) O'Brien-Fleming type spending function and a futility boundary derived from the γ-spending function of Hwang et al. 1990 with parameter $\gamma = -5$. It was decided, with the help of the East (2008) software, to keep the study open for a maximum of 732 OS events, with one interim analysis after 440 events (60% of the total information), whereby a 1-sided level-0.025 group sequential logrank test would have 85% power to detect a hazard ratio of 0.8. As this was an event-driven trial, sample size did not play a direct role in the above power calculation. Nevertheless, the rate of accrual, duration of accrual, and duration of follow-up would affect the total study duration or time needed to obtain 732 events. Again, with the help of East, it was determined that by enrolling 950 subjects over a 2-year period and following them for an additional 7 months, the required 732 OS events would be expected to arrive by the end of the follow-up period.

Now, the assumption of 7.4 months for median survival on the control arm was based on published results from a previously completed large, well-controlled trial. It was felt, however, that due to improvements in standard of care, the median survival might now be as high as 7.8 months. If that were the case, the underlying hazard ratio would be 0.84, and the power would drop to about 65%. For this reason, it was decided to permit an adaptive increase in the number of events up to 50% (from 732 to 1,098), if the interim results fell into the promising zone, here defined as conditional power between 50 and 85%. The magnitude of the increase in events was to be the amount needed to recover 85% conditional power. It was decided that the sample size would be increased in the same ratio as the increase in events.

For time to event trials, the test statistic following an adaptive increase in the number of events has a different form than that displayed in (4.3). This is because some of the patients enrolled in the first cohort will still be censored at the time of the interim analysis. To properly account for the data from the two stages, we rely on the fact that the logrank statistic, if properly standardized, has independent increments. To be specific, let $D_{\max}$ be the maximum number of events required to achieve $1-\beta$ power. Let, D_1 denote the number of events at the time of the interim analysis. Let, Z_1 be the logrank statistic observed at the interim look. Let, Z_2 be the logrank statistic we would observe at the final analysis if there were no adaptive change in the maximum number of events. Now suppose that, based on the observed $Z_1 = z_1$, the maximum number of events is increased from $D_{\max}$ to $D^*_{\max}$. Let, Z^*_2 be the logrank statistic at the final analysis, where the number of events is $D^*_{\max}$. Define

Table 4.11 Operating characteristics of classical and adaptive group sequential designs

	Classical group sequential		Adaptive group sequential	
Hazard ratio	Power (%)	Expected events	Power (%)	Expected events
0.80	85	617	87	656
0.81	81	630	83	670
0.82	76	641	78	684
0.83	71	649	73	693
0.84	64	657	68	701

All Plan 4 results are based on 100,000 simulated trials

the information fractions $t_1 = D_1/D_{\max}$ and $t_2^* = D_{\max}^*/D_{\max}$. To preserve the type-1 error of the test for H_0, we must use the weighted statistic

$$T_2 = \sqrt{t_1}Z_1 + \sqrt{1-t_1}\left\{\frac{\sqrt{t_2^*}Z_2^* - \sqrt{t_1}Z_1}{\sqrt{t_2^* - t_1}}\right\}. \tag{4.7}$$

Notice that the weights for the two cohorts are based on the *prespecified* events D_1 and $D_{\max}$. The formula for conditional power at the interim analysis is likewise different from (4.6), reflecting the fact that some subjects enrolled in the first cohort are still censored at the time of the interim analysis. The conditional power formula is

$$\mathrm{CP}_{\hat{\delta}_1}(z_1) = 1 - \Phi\left[\frac{z_\alpha\sqrt{D_{\max}} - z_1\sqrt{D_1} - \hat{\delta}_1\sqrt{r(1-r)}\sqrt{D_{\max}^* - D_1}\sqrt{D_{\max} - D_1}}{\sqrt{D_{\max} - D_1}}\right], \tag{4.8}$$

where $\hat{\delta}_1$ is the estimate of the log hazard ratio at the interim, and r is the randomization fraction (here equal to 0.5). For additional details concerning the derivation of (4.7) and (4.8) refer to Wassmer (2006).

The operating characteristics of the lung cancer trial are displayed in Tables 4.11 and 4.12 for underlying hazard ratios between 0.8 and 0.84. Table 4.11 displays the overall operating characteristics while Table 4.12 displays them by zone. In each table, the classical group sequential design and the adaptive group sequential design are compared.

The results follow a similar pattern to what was observed in the previous two examples. Table 4.11 shows that the adaptive design produces a very modest gain in terms of overall power, ranging from 2% at a hazard ratio of 0.8–4% at a hazard ratio of 0.84. The cost in terms of additional expected events is likewise modest. Table 4.12, however, shows that there is a gain of between 10 and 13% power if the interim outcome falls in the promising zone. This is the appeal of the adaptive design. Consider, for instance, the prospects for a successful trial if the true hazard ratio is 0.84. The interim outcome will fall in the unfavorable zone 38% of the time, in which case the prospects for a successful trial are equally bleak for both the

Table 4.12 Operating characteristics of Plan 1 and Plan 4 conditional on interim outcome

Hazard ratio	Interim outcome	Probability of interim outcome (%)	Power conditional on interim outcome		Expected events	
			Classical (%)	Adaptive (%)	Classical	Adaptive
0.80	Unfavorable	21	49	49	705	706
	Promising	17	83	94	732	951
	Favorable	62	98	98	556	556
0.81	Unfavorable	25	44	44	701	701
	Promising	18	80	92	732	952
	Favorable	57	07	97	565	565
0.82	Unfavorable	29	39	39	697	697
	Promising	19	77	90	732	955
	Favorable	51	96	96	575	575
0.83	Unfavorable	34	35	35	692	691
	Promising	20	74	87	732	957
	Favorable	46	95	95	583	583
0.84	Unfavorable	38	30	30	687	686
	Promising	20	71	84	732	958
	Favorable	42	94	94	591	591

All results are based on 100,000 simulated trials

classical and adaptive designs, but so are the expected number of events. The interim outcome will fall in the favorable zone 42% of the time, in which case the prospects are excellent for both the classical and adaptive designs, and so are the expected number of events. The remaining 20% of the time, the interim outcome will fall in the promising zone and this is where the adaptive design will help by boosting up the power from 71 to 84%. The expected number of events, hence, other related resources like sample size and study duration will also increased in the promising zone. Presumably the power gain justifies the use of these additional resources.

5 Concluding Remarks

The statistical methodology that permits sample size reestimation based on an interim estimate of the treatment effect has been available for about 10 years and has gradually made its way into actual confirmatory clinical trials. All three examples discussed in the current paper are based on real trials whose designs were accepted by the FDA, and all have been activated subsequently. The recently released FDA Draft Guidance for Industry on Adaptive Design (2010) is an indispensible document for sponsors who are considering the adaptive route for their confirmatory trials. It classifies adaptive approaches into those that are generally well-understood and those that are less well understood. Group sequential designs and blinded sample size reestimation fall into the category of well-understood methods while sample size reestimation using unblinded estimates of the treatment effect falls into

the category of less well understood methods. This is not to suggest that the latter methods are disallowed. Rather, the guidance document says that these methods should be used when other better understood methods are unable to meet the primary study objectives. In our context, this implies that one should be able to provide a valid reason for including the option for unblinded sample size reestimation at an interim look in preference to using a conventional design that fixes the maximum sample size at the outset. The usual reason is that even after a thorough review of all relevant prior data, there is still some uncertainty regarding the true treatment effect and population variability. Sometimes a group sequential design powered to detect a small but clinically meaningful treatment effect will resolve this difficulty. We have seen, however, that overruns, large up-front commitments, and conservative boundaries for early efficacy stopping are often deterrents to the group sequential approach.

The reason for the FDA's classification of unblinded sample size reestimation into the "less well understood" category has nothing to do with the statistics. The validity of the statistical methodology is accepted. The FDA's real concern is the possibility of operational bias. Operational bias might be introduced into the study if the interim results were somehow revealed to the investigators and led to selective withdrawal of patients from the study before they had completed their full course of treatment. The only way to prevent this type of operational bias is by creating good operating procedures that are built into an interim analysis charter and are strictly followed. Such procedures are already in place for group sequential designs, and can be modified for the specifics of the adaptive setting. To this end, it is customary to set up an independent statistical center (ISC) for creating the interim analysis report and an independent interim analysis review committee (IARC) whose task it is to review the interim analysis report and implement the sample size reassessment in accordance with the interim analysis charter. If the study already has a functioning data monitoring committee (DMC), that committee or a subset of that committee could fulfill the role of the IARC. The IARC charter would, however, differ from that of a traditional DMC. This charter should describe how the interim data are to be transferred to the ISC, what the interim analysis report should contain, who may have access to that report, the precise rules for altering the sample size, and the procedure to be followed in making the sample size recommendation to the sponsor. Access to the charter must be restricted to the ISC, the IARC, and only those employees of the sponsor organization who were involved in the trial design. Everyone with access to the charter should be required to sign a confidentiality agreement in order that the precise rules governing sample size reestimation may not be disclosed to the outside community.

Even with good operating procedures in place to prevent premature disclosure of interim results, the mere fact that a sample size increase was implemented cannot be hidden from the sites enrolling subjects. Concerns have been raised that this knowledge alone could modify investigator behavior. It is difficult, however, to anticipate how investigators would interpret this knowledge. Some might feel that the chances of the trial succeeding have improved while others might take it as a sign that the initial estimate of treatment effect was too optimistic. A discussion about

the rationale for the adaptive design, its potential for reducing the risk of running an underpowered study, and the need to maintain the same pattern of patient enrollment throughout the study might be an important agenda item at an investigator meeting. It would likewise be important to educate the Institutional Review Boards about the adaptive nature of the design, stressing that while the actual sample size is not known at the outset, the maximum sample size if an adaptation takes place has been fixed. In this sense, the uncertainty about the sample size of an adaptive design is similar to the uncertainty about the sample size of a group sequential design or of any event driven trial in which the number of events but not the sample size is known in advance.

Additional discussion concerning these operational issues is provided in a White Paper published by the PhRMA Adaptive Working Group (2007). The FDA has adopted a "wait-and-see" attitude to gain more experience with the risks and benefits of unblinded sample size reestimation. However, a well prepared submission that addresses both the statistical and operational issues and backs them up with simulations and a detailed charter stands an excellent chance of regulatory acceptance.

In the early years following publication of methods for unblinded sample size reestimation, concerns were raised about their efficiency relative to conventional group sequential methods. These concerns arose because the type-1 error of an adaptive design can only be preserved by use of a weighted statistic like (4.3) with prespecified weights. This violates the sufficiency principle. Tsiatis and Mehta (2003) demonstrated that for any adaptive design with sample size modification requiring the use of the weighted statistic (4.3), one could construct a group sequential design utilizing the usual sufficient statistic that would stop earlier with higher probability of rejecting the null hypothesis if $\delta > 0$ and also stop earlier with higher probability of accepting the null hypothesis if $\delta < 0$. Jennison and Turnbull (2003) demonstrated a similar result empirically by creating a group sequential design with the same power function as an adaptive design and then demonstrating by simulation that it would have a smaller expected sample size. These results are of great theoretical interest but of limited practical value for sponsors of industry trials. They produce appreciable efficiency gains only if there are no overruns, a large number of interim analyses, a large up-front sample size commitment and aggressive early-stopping boundaries. Sponsors are usually unwilling or unable to impose these conditions on their trial designs. Furthermore, the analyses of Tsiatis and Mehta (2003) and Jennison and Turnbull (2003) do not capture the essential appeal of the adaptive approach which is to invest limited resources initially and to invest more only after seeing interim results that are promising. Most adaptive designs have only two stages, increase the sample by at most a factor of 2, and that too, only if the results fall in a promising zone. It has yet to be demonstrated that there is any appreciable loss of efficiency due to the use of the weighted statistic in these settings.

When the Tsiatis and Mehta (2003) paper was reviewed, a referee raised an additional objection to the use of the weighted statistic (4.3). It was pointed out that if the sample size is increased following an interim analysis, then the contribution

of subjects in the second cohort will be down-weighted relative to the subjects in the first cohort. This is evident from (4.3) where, although the sample size of the second cohort is $n^{(2*)}$, the weights combining the data from the two stages use $n^{(2)}$. Conversely, if the sample size of the second cohort is reduced, then the data from the first cohort will be down-weighted. This type of down-weighting would appear to be the price one must pay for altering the sample size in a data dependent manner. It turns out, however, that the price is rather small, and in some cases nonexistent. A paper by Chen et al. (2004) proved that if the sample size is only increased when the conditional power $\mathrm{CP}_{\hat{\delta}_1}(z_1)$ is at least 50%, it is not necessary to use the weighted statistic. One can use the conventional statistic and the type-1 error will be preserved notwithstanding the data dependent sample size increase. Subsequently, Gao et al. (2008) and Mehta and Pocock (2010) were able to relax this condition so that the promising zone may begin at conditional power values as low as 30% depending on the magnitude of the sample size increase. We have observed, through simulations in East (2008), that for the three examples in the current paper that there is no inflation of type-1 error if the conventional statistic is used following an adaptive sample size increase. Indeed, the promising zone for the lung cancer example starts at 50% conditional power, and therefore satisfies the Chen et al. (2004) criterion.

It has been argued by Bauer and Koenig (2006) that the use of conditional power based on $\hat{\delta}_1$ should be avoided because the estimate lacks precision. Our examples have shown, however, that it works well in practice. Recall that the goal at the interim analysis is not to estimate δ with great precision but to merely ascertain whether the current results show promise. Thus, an interim estimate that falls inside the promising zone suggests that the initial choice of δ might have been a bit optimistic, and now is the time to improve the chances of a successful outcome by increasing the sample size to detect a more modest but still clinically meaningful true effect, if it does exist. One can, if preferred, represent the promising zone in terms of $z_1, \hat{\delta}_1$, or $\mathrm{CP}_{\hat{\delta}_1}$, since these statistics are all transformations of one another.

We have focused throughout this article on situations in which the sample size is increased. In fact, the methods discussed here will preserve the type-1 error for either a sample size increase or a sample size decrease. However, the FDA Adaptive Guidance Document (2010) makes it clear that from a regulatory perspective, the use of adaptive designs for decreasing the sample size will not be permitted. In its place, they advocate the use of group sequential designs with futility boundaries.

Although the focus of this article has been clinical trials sponsored by pharmaceutical companies, these designs could be equally useful in government sponsored trials run by the cooperative groups. Moreover, the cooperative groups have the flexibility to either expand or reduce sample sizes, since the participating treatment sites are on long-term grants and are associated with a single coordinating center.

An important issue that has not been discussed at all is how to provide a valid point estimate and confidence interval for the combined data in which an adaptive sample size change was made. This is a more difficult problem and research is still on-going. We refer the reader to articles by Mehta et al. (2007), Brannath et al. (2009) for recent results.

In this article, we have discussed both the motivation and statistical methodology for unblinded sample size reestimation in confirmatory trials. We have shown through real examples, that a major benefit of this type of adaptive design is that it reduces the risk of running an underpowered study without a large initial investment. The two-stage nature of the investment, with the second installment being obligated only if the interim results have significantly increased the odds of success, often makes the adaptive design more attractive than a conventional design. A major additional benefit of the adaptive approach is flexibility. The adaptive methodology controls the type-1 error even if the prespecified criteria for increasing the sample size are overruled at the interim analysis. This might be desirable for a variety of reasons both internal and external to the current trial. For example, in addition to observing a promising outcome at the interim analysis, the safety profile for the test drug might turn out to be far superior to what was originally anticipated, and this might make the new drug more competitive in the marketplace. One could, therefore, justify increasing the sample size by a larger amount than that determined by the prespecified rules, and thereby further reduce the chances of a false negative outcome. Another possible situation in which one might overrule the prespecified criteria for sample size change would be if compelling results from other clinical trials on comparable populations, treated with the same class of drugs became available and caused the sponsor to revise the value of δ at which to power the current study. Ideally, one would wish to adhere strictly to the prespecified criteria for sample size change since the operating characteristics of the design would change if they were overruled. This would certainly be the preference of regulatory authorities. As a practical matter, however, it is not possible to anticipate every contingency under which a sample size change is desirable. It is a strength of the adaptive approach that the validity of the statistical test at the end of the trial is not affected by unanticipated developments arising over the course of the clinical trial that necessitate making changes to the prespecified criteria for sample size adaptation.

Finally, for confirmatory trials, regulatory approval must be secured in advance through a special protocol assessment (SPA) or an end-of-phase-2 meeting. For this purpose, the sponsor is required to submit the protocol, the charter, and the simulations backing up the statistical validity of the proposed adaptive approach. The charter should address all logistical and operational issues. An adaptive design might not always be the right choice. The more established fixed sample and group sequential designs should always be evaluated alongside an adaptive design. Simulations play a crucial role in understanding the operating characteristics of an adaptive design and deciding whether it is an appropriate choice for the trial under consideration. There should be a tangible, quantifiable benefit arising from the decision to take the adaptive route.

Acknowledgements The author thanks Dr. Howie Golub for helpful discussions on adaptive designs and Dave Harrington for critical comments that have greatly improved the article.

Software support for this article was provided by the East (2008) software package developed by Cytel Corporation.

References

Bauer P, Koenig F (2006) The reassessment of trial perspectives from interim data – a critical view. Stat Med 25(1):23–36

Bauer P, Kohne K (1994) Evaluation of experiments with adaptive interim analyses. Biometrics 50(4):1029–1041

Brannath W, Mehta CR, Posch M (2009) Exact confidence bounds following adaptive group sequential tests. Biometrics 65(2):539–546

Chen YH, DeMets DL, Lan KK (2004) Increasing the sample size when the unblinded interim result is promising. Stat Med 23(7):1023–1038

Cui L, Hung HM, Wang SJ (1999) Modification of sample size in group sequential clinical trials. Biometrics 55(3):853–857

East (2008) Software for design and monitoring of group sequential and adaptive trials; version 5.3. Cytel Inc., MA. www.cytel.com

FDA (2004) Innovation or stagnation: Challenge and opportunity on the critical path to new medical products. http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html

FDA (2010) Guidance for industry: Adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf

Gao P, Ware JH, Mehta C (2008) Sample size re-estimation for adaptive sequential design in clinical trials. J Biopharm Stat 18(6):1184–1196

Haybittle JL (1971) Repeated assessment of results in clinical trials of cancer treatment. Br J Radiol 44(526):793–797

Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. Academic, New York

Hwang IK, Shih WJ, DeCani JS (1990) Group sequential designs using a family of type i error probability spending functions. Stat Med 9:1439–1445

Jennison C, Turnbull BW (2003) Mid-course sample size modification in clinical trials based on the observed treatment effect. Stat Med 22(6):971–993

Lan KKG, Demets DL (1983) Discrete sequential boundaries for clinical trials. Biometrika 70(3):659–663

Lehmacher W, Wassmer G (1999) Adaptive sample size calculations in group sequential trials. Biometrics 55(4):1286–1290

Mehta C, Gao P, Bhatt DL, Harrington RA, Skerjanec S, Ware JH (2009) Optimizing trial design: Sequential, adaptive, and enrichment strategies. Circulation 119(4):597–605

Mehta CR, Pocock SJ (2010) Adaptive increase in sample size when interim results are promising: A practical guide with examples. Stat Med, Wiley Online Library, November 2010

Mehta CR, Bauer P, Posch M, Brannath W (2007) Repeated confidence intervals for adaptive group sequential trials. Stat Med 26(30):5422–5433

Muller HH, Schafer H (2001) Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. Biometrics 57(3):886–891

PhRMA (2007) White paper of the phrma adaptive working group. DIA J

Pocock SJ (2005) When (not) to stop a clinical trial for benefit. JAMA 294:2228–2230

Proschan MA, Hunsberger SA (1995) Designed extension of studies based on conditional power. Biometrics 51(4):1315–1324

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

Wassmer G (2006) Planning and analyzing adaptive group sequential survival trials. Biom J 48(4):714–729

Woodcock J, Woosley R (2008) The fda critical path initiative and its influence on new drug development. Annu Rev Med 59:1–12

Chapter 5
On Stopping a Randomized Clinical Trial for Futility

Jay Herson, Marc Buyse, and Janet Turk Wittes

1 Introduction

Many modern Phase 3 clinical trials incorporate formal planned interim analyses. For reasons of ethics and efficiency, sophisticated stopping rules, based on accumulating efficacy data, allow trial sponsors to take various actions at interim time points under the stewardship of an independent data monitoring committee (DMC) (Ellenberg et al. 2002; US Food and Drug Administration 2006; International Conference on Harmonisation 1998; Herson 2009). At these interim times the sponsor may terminate the trial because the accumulated data have demonstrated superiority or inferiority of an experimental treatment over a control. The investigators or sponsor might also elect to terminate the trial because of futility, that is, because the interim data imply a very low likelihood of observing statistically significant superior efficacy if the trial continues to termination, or the anticipated superiority at the end of the trial is deemed disappointing. They may specify an effect size that is not of clinical interest and define a trial to be futile if it clearly shows that the effect of the new treatment is less than that size. In this case, the final report from a trial that stops early can include an explicit statement of the effect size the trial has ruled out.

J. Herson (✉)
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
e-mail: jay.herson@earthlink.net

M. Buyse
IDDI Consultants, International Drug Development Institute, Louvain-la-Neuve, Belgium
Department of Biostatistics, Hasselt University, Diepenbeek, Belgium
e-mail: marc.buyse@iddi.com

J.T. Wittes
Statistics Collaborative, Inc., Washington, DC 20036, USA
e-mail: janet@statcollab.com

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7_5,

Another purpose of interim analysis may be to recalculate the sample size so at the end of the trial the power will be adequate to reach statistical significance for some effect size of interest (perhaps not the original protocol-specified effect size) at the end of the trial. This chapter will focus on testing practices for futility; other chapters in this book cover sample size recalculations and interim analyses for efficacy. Chuang-Stein et al. (2006) review methods of sample size recalculation.

Clinical trialists involved in trials sponsored by government agencies such as the United States National Institutes of Health (NIH) or the British Medical Research Council (MRC) in nonregulatory settings and by industry in globally-regulated settings, have accrued sufficient experience with futility methods to allow rethinking their role and use. Sponsors, regulators, investigators, investors, and DMC members now often request inclusion of a formal futility boundary as part of the interim analysis of confirmatory trials. Many recent trials have either terminated early for futility or have had protocols with formal futility stopping rules (see Table 5.1). Futility analyses do not often appear in exploratory trials but Lee and Feng (2005) have made the case for including futility stopping rules in randomized Phase 2 protocols in oncology. Emerson et al. (2007) provide a comprehensive discussion of sequential designs that shows the relationship between sequential analysis and stochastic curtailing.

This paper presents the rationale and current methods for futility analysis, describes their advantages and disadvantages in various settings, and provides suggestions for implementing – or not implementing – futility guidelines. We argue that while futility boundaries may sometimes be useful, they may lead to premature termination of a trial, and therefore to equivocal results. In particular, we caution against stopping for futility in settings where early stopping of a trial is likely to miss a late effect of treatment, either beneficial for a new product or harmful for a product already in use. We encourage the use of confidence intervals for reporting the estimated effect sizes in studies that stop for futility so that clinicians can interpret the data in terms of the size of effect that the study has ruled out. While this point may be obvious to statisticians, many reports of trials that stop for futility fail to report the results fully.

2 Inclusion of a Futility Boundary in a Protocol

2.1 Trial and Sponsor Characteristics Favorable to Inclusion

Several considerations may justify including futility analyses in clinical trials. First, exposing patients to the potential side effects of a therapy in a clinical trial that has little chance of demonstrating efficacy may be undesirable, or even unethical. The decision to stop for futility would usually emerge from a planned interim analysis using statistically justified stopping rules that provide a guideline for futility (Proschan et al. 2006). Futility guidelines are often useful in trials where harms

Table 5.1 Recent examples of trials either terminated early for futility or having protocols with formal futility stopping rules

Disease	Reference	Method of futility analysis
Cardiovascular disease	Teerlink et al. (2005)	Stochastic curtailing and conditional power calculations
	Mas et al. (2006)	Safety considerations and conditional power
Pulmonary disease	US National Heart, Lung and Blood ARDS Clinical Trials Network(2004)	Group sequential boundary
	Curley et al. (2005)	Group sequential boundary
	Robbins et al. (2008)	Unplanned conditional power analysis
Orthopedics	Kiel et al. (2007)	Conditional power and beta spending function
Neurology	Gorelick et al. (2003)	Conditional power
Infectious disease	Abraham et al. (2005)	Conditional power
	Abraham et al. (2003)	No formal method specified
Hematology	Bussel et al. (2007)	Group sequential boundary
Kidney disease	Singh et al. (2006)	Group sequential boundary
Oncology	Geyer et al. (2006)	Group sequential boundary
	Ajani et al. (2008)	Group sequential boundary
	Stadler et al. (2005)	Bayesian method
	Spriggs et al. (2007)	Stochastic curtailment
	Gennari et al. (2006)	Bayesian, not preplanned
	Lanciano et al. (2005)	Planned but based on unfavorable results in experimental group, no formal statistical method used
	Lorigan et al. (2007)	Planned based on interim results, no formal statistical methods used

are likely (e.g., anticancer, antidiabetes, cox-2 inhibitors). A second reason is that inclusion of a nonbinding futility guideline can motivate complete DMC discussion of the importance of observed safety differences between treatments when the futility criteria are met. Without a prespecified guideline, DMC members may feel that the treatment difference in safety profile is acceptable or expected and they may see no reason to defend a decision to continue the trial. Trials with short-term outcomes lend themselves to futility boundaries because delayed effects of therapy would not be encountered. Examples would be 3-day pain outcome for analgesics, 30-day mortality in sepsis, or a 14-day anti-infective trial.

The typical oncology add-on trial that compares treatment A to A + B, where A is an active control and B is an experimental therapy, also lends itself to a futility guideline because even in the absence of a toxicity burden by treatment B, a two-drug regimen is difficult to justify.

Another reason to include a futility guideline might simply be financial: a pharmaceutical company may be reluctant to continue to commit large investments in a therapy that is likely to be less beneficial either to patients, to the company, or to both, than originally anticipated. Similar conditions may apply to government-sponsored trials – once the likely results of a trial are apparent to a DMC, expenditure of public monies to address other questions might be preferable to continuing the trial. In cases where a commercial sponsor has an active pipeline of drug candidates ready for testing futility analyses might add efficiency to drug development.

2.2 Trial and Sponsor Characteristics Less Favorable to Inclusion

In several situations, futility guidelines are less desirable (see Table 5.2). For example, if the intervention is in current use, such as in estrogen–progestin hormone replacement therapy, and the DMC judges that stopping for futility would be unlikely to change practice (Wittes et al. 2007), futility guidelines may not be useful. A controversial or novel intervention (e.g., genomic treatment selection in oncology, drug-eluting stents, or even a trial that compares an oral and injectable formulation of the same product) might require a complete trial to be persuasive.

Another reason for not including a futility boundary would be in trials studying time to event when the effect of the intervention may take years to manifest itself. Examples of interventions with late benefits are lipid-altering in coronary disease (Robins et al. 2001), and immunotherapies in cancer (Loughlin 2008; Gray et al. 2008). Trials that stop because the observed effect size is very small may miss long-term effects.

Trials with time-to-event outcomes and rapid enrollment, for example, those with a short treatment phase and a long follow-up such as in vaccine trials, do not lend themselves to futility stopping. Here, at the time of an interim analysis perhaps all, or almost all, patients will have been enrolled so that termination for futility has little utility.

Some trials, like those involving implantable medical devices (e.g., vagus nerve stimulators for epilepsy and depression, mechanical and biosprosthetic heart valves), may have a learning curve for investigators rendering a futility analysis inappropriate.

Table 5.2 Inclusion of a futility guideline in a protocol

a. Trial and sponsor characteristics favorable to including futility guidelines

- Issues related to safety are anticipated
- To motivate complete DMC discussion of borderline issues related to safety when a futility guideline is crossed
- Trials with short term outcomes
- Trials using add-on design comparing A vs. $A + B$ where A is active control and B is experimental treatment
- Financial reasons – sponsor may not want to continue to invest in a trial not showing obvious benefit at interim
- A full pipeline of drugs is ready for testing

b. Trial and sponsor characteristics less favorable to including futility guidelines

- Intervention is in current use and persuasive evidence would be needed to change practice
- A trial for a novel intervention would likely need a complete trial to be persuasive to clinicians and regulators
- Time to event outcomes when effect of treatment may take years to manifest
- Time to event outcomes and rapid enrollment, such as vaccine trials, because all patients are likely to be enrolled at the time of interim analysis
- Trials that may involve a learning curve for investigators (e.g., surgical techniques involved with implantable medical devices)

3 Factors Influencing a Decision to Terminate for Futility

While early termination of a clinical trial for superiority is relatively straightforward, futility analysis is complicated because, even when the efficacy data from the trial have reached a futility boundary, a DMC may be reluctant to recommend early stopping or the investigators (or sponsor) may choose not to stop the trial. Statistical methods for futility stopping may be "binding" or "nonbinding" with the latter appearing most frequently in protocols. When the futility guideline is nonbinding, the DMC will typically consider a range of factors in making a decision for recommending termination. DMCs may also choose to overrule terminating for superiority but many DMCs feel more pressure, ethically and scientifically, to terminate in that case than in futility analysis because when futility is indicated, unless the safety profiles differ, the data are consistent with both treatments groups' providing similar benefit.

Below, we now summarize some reasons for deciding whether or not to recommend terminating a trial when futility guidelines are reached. In each case we distinguish situations where the DMC has the tools to make a decision on its own and those in which the DMC will need some dialog with the investigators or a sponsor representative not involved with the trial to aid in making its decision.

3.1 Reasons to Terminate for Futility

Several factors might lead a DMC to recommend terminating a trial when a futility guideline is reached. The most obvious is a treatment difference in safety large enough to make continuing the trial unethical. Also, secondary endpoints may show consistent and reliable trends against the experimental treatment. External information may become available showing that similar interventions (e.g., drugs of the same therapeutic class) fail to show efficacy for this indication. Conversely, external information may show that competing drugs showing larger treatment effects than planned or expected from this trial have raised the efficacy bar for this indication. In such a case, a lack of effect in the present trial would probably not call for early termination, but would raise a red flag as to whether the new drug or intervention was applied properly.

Other reasons to terminate would require input from sponsor representatives that might not be known to DMC members. Financial considerations might make it imprudent to continue on this path. If a sufficient pipeline of drugs is available for testing, the sponsor may wish to move on. Even a seemingly unimportant benefit-risk differential might cause a sponsor to terminate development of that product or that indication. For example, a trial for seasonal allergic rhinitis might reach a futility boundary but the only safety difference is that the experimental treatment has a 10% incidence of transient headache while the active control group has no cases and other competing treatments do not have this side effect. The sponsor may see an obvious marketing burden with this product and wish to terminate. Once again because these reasons to terminate involve business decisions the DMC will have to consult with a sponsor representative not involved with the trial such as a pharmacovigilance officer or the steering committee.

3.2 Reasons Not to Terminate for Futility

Several reasons may make a DMC reluctant to recommend terminating for futility. First, the data may exhibit time trends. For example, eligibility requirements may have changed during the trial, where an initiative was started to enroll more patients of a certain type such as African-American patients or those in a diagnostic subgroup such as stage 2 patients in a trial for stage 2/3 melanoma patients or where previously unrepresented patient types are now enrolling because competing trials have ended. Levy (2004) argued against stopping for futility in the ARDS trial (US National Heart, Lung and Blood Institute ARDS Clinical Trials Network 2004) where he noticed that treatment groups were unbalanced at baseline for certain prognostic factors. In addition, the ARDS trial used statistically sophisticated methods to impute missing data. Some DMCs might be uncomfortable with recommending terminating a trial where futility was determined on the basis of such imputed data. In some trials, data quality may be suboptimal and or a smaller than desirable number of centers may have had site monitoring. Secondary

outcomes might be showing benefit and external information on similar therapies may have shown beneficial results for this indication. For a time-to-event endpoint, such as time to progression, if all or almost all patients are already enrolled in the trial, the DMC may feel it prudent simply to wait for the final analysis before making a recommendation to terminate. In trials using a presumed surrogate endpoint for the interim analysis of futility but not in the final analysis (e.g., progression-free survival used for futility at interim and overall survival for the final analysis), DMCs may be reluctant to recommend termination if the members feel that the two endpoints have low correlation, as has been demonstrated for progression-free and overall survival in advanced prostate cancer (Sternberg et al. 2009). In industry-sponsored trials, when futility boundaries are crossed consultation with a sponsor representative might yield additional reasons not to terminate. The sponsor may feel that a complete trial might be useful to gain approval in other countries. Continuing the trial might allow the sponsor to learn more about response to treatment in important subgroups and allow collection of information vital for planning new trials. See Table 5.3 for some factors involved in a decision to terminate for futility.

As described in the following section, the statistical method most appropriate to assess futility during the trial may also depend, in part, on the characteristics listed in Tables 5.2 and 5.3.

4 Methods of Futility Analysis

Statistical methodology for futility analysis has at least three flavors. One, which has both classical and Bayesian varieties, is based on stochastic curtailment; a second is based on group sequential analysis; a third uses an outcome different from the primary outcome to assess futility.

Deterministic curtailment, introduced by Alling (1963, 1966), is the process of ending a trial early when no matter whatever the next observations may be, the result in terms of rejection or nonrejection of the null hypothesis cannot change. For example, suppose an experiment with a planned sample size of N requires r successes to reject the null hypothesis. If n observations have been made and r successes have already occurred, then no matter what the next $(N - n)$ observations are, the null hypothesis will be rejected. Thus, stopping the trial at n leads to the certain conclusion that the null hypothesis will be rejected. (Of course, the observed proportion of successes is not an unbiased estimate of the underlying proportion.) Lan et al. (1982) extended deterministic curtailment by adding a probabilistic element to Alling's method – stochastic curtailment is stopping a trial early with the ability to assign a probability of rejecting or not rejecting the null hypothesis. Although philosophically different, conditional and predictive power, which is discussed below, both belong to the class of methods of stochastic curtailment. Both seek to compute a probability of rejecting the null hypothesis at the end of the trial conditional on accumulated data and certain assumptions. The result of a conditional or predictive power calculation is a probability of a statistically significant result at the end of the trial.

Table 5.3 Factors influencing a decision to terminate for futility

a. Reasons to terminate for futility

Decisions to terminate that can usually be made by DMC independent of sponsor

- A disturbing difference in safety profile is observed
- Secondary endpoints show consistent and reliable trends against experimental treatment
- External information arises on lack of efficacy of similar therapies for this indication
- The efficacy bar has been raised as external information becomes available indicating that competing therapies are achieving effect sizes larger than those planned or expected in this trial

Decisions to terminate usually requiring DMC interaction with sponsor representative

- Financial reasons not to continue to invest in this drug
- Sponsor has another promising therapeutic candidate waiting
- Benefits and risks not favorable for modest toxicity – the drug has side effects that while not serious are annoying to the patient and are not present in other drugs for this indication on the market

b. Reasons not to terminate for futility

Decisions not to terminate that can usually be made by DMC independent of sponsor

- Patient characteristics change over the course of the trial. Change may be due to factors such as modifications in eligibility requirements, a decision to target patients of different types such as African-Americans or Stage 2 patients, sicker patients entering early in the trial, different centers entering the study at different times
- Baseline imbalances in prognostic factors
- Data may be of questionable quality
- Site monitoring may be insufficient
- Primary efficacy endpoint requires considerable imputation of missing data
- Secondary outcomes are suggestive of patient benefit from experimental treatment
- External information arrives on benefit of similar therapies for this indication
- The outcome measures some time-to-event and all patients are enrolled
- The interim analysis of futility is based on a presumed surrogate outcome for the primary efficacy outcome that will be analyzed. The DMC members do not believe these endpoints to be sufficiently correlated.

Decisions not to terminate usually requiring DMC interaction with sponsor representative

- Sponsor feels the data can be used to gain approval in other countries
- Sponsor staff want to learn more about treatment effect in certain subgroups of patients
- Sponsor staff can use the data from a complete trial to plan new trials

Group sequential futility methods use some of the Type II error rate to test the alternative hypothesis in a manner analogous to group sequential efficacy methods that use some Type I error rate to test the null hypothesis. They specify a priori an effect size too small to be of clinical interest and then create a group sequential boundary such that, if crossed, the data have demonstrated – with, of course, some use of Type II error rate – that the true effect is less than the size deemed to be of no clinical interest.

The third approach we discuss is completely different in spirit – it bases its conclusion on a biomarker believed to be highly predictive of the outcome of interest. A Phase 3 trial studying whether or not the intervention is effective on a long-term clinical outcome may stop early if the effect on a intermediate endpoint is too weak to be consistent with a clinically important effect on the outcome of interest, whether or not the intermediate outcome is a good surrogate for the long-term endpoint.

The following sections describe these methods in more detail. We shall assume throughout this chapter that the hypothesis of interest is one of treatment difference, rather than of equivalence or noninferiority of the experimental treatment as compared with the control group. For an example of a trial that addressed futility in the context of a noninferiority hypothesis, see Cairns et al. (2008).

4.1 Conditional Power

Techniques based on conditional power consider projecting the outcome of the trial conditional on the data observed thus far and incorporating some assumptions concerning the true effect size that will operate during the remainder of the trial.

To calculate conditional power, we start with the *B*-value (Lan and Wittes 1988) $B(t)$, where t is the information time of the study ($0 \leq t \leq 1$), $z(t)$ is the observed z-value at time t, and $B(t) = z(t)t^{1/2}$. Let θ, which is called the "drift parameter," denote the expected value of the z-score at the end of the trial; that is, $E\{B(1)\} = \theta$. Note that $z(t) = \hat{\mu}(t)/SE[\hat{\mu}(t)] = t^{1/2}\theta$. Therefore, $E\{B(t)\} = \theta t$ or $\theta = E\{B(t)\}/t$.

Thus, $B(1) = z(1)$; in other words, the *B*-value at the end of the trial is identical to the z-value at the end of the trial. Under a constant effect size θ, or drift, $B(t)$ increases linearly but $z(t)$ does not so that the *B*-value is a more natural statistic to monitor than is the z-value.

As Proschan et al. (2006) describe, the conditional mean of $B(1)$ given $B(t) = b$ has a simple geometric interpretation. If the drift size is θ (and constant over the life of the trial), then the conditional mean of $B(1)$ can be depicted as the endpoint of a line segment that starts at (t, b) and extends to the end of the trial with slope θ (see Fig. 5.1). Superimposing on this conditional expectation a normal distribution with mean $b + \theta(1-t)$ and variance $(1-t)$ on its side, conditional power is the area above the point $(1, z_{\alpha/2})$. In other words, conditional power is the probability that the study will show statistical significance if it continues to its planned end and if the future effect size θ is constant.

In general, if the drift is constant over time, the conditional power at time t is:

$$\mathrm{CP}(t) = 1 - \Phi[(z_{\alpha/2} - \{b + \theta(1-t)\})/(1-t)^{1/2}]. \tag{5.1}$$

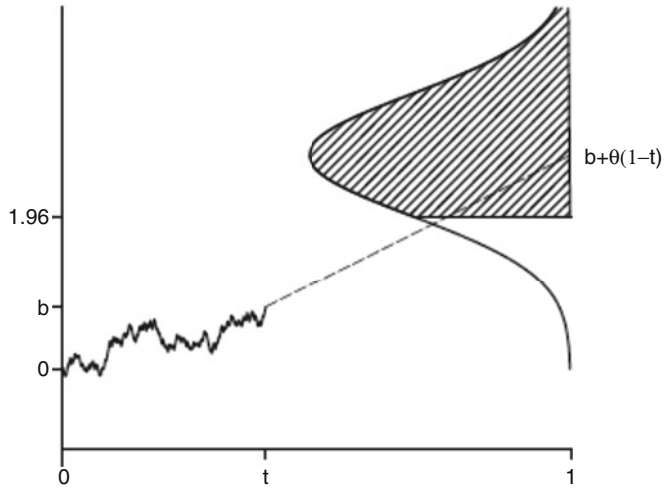

Fig. 5.1 Given a trial with data through information time t and a B-value $B(t) = b$, the B-value at the end of the trial, $B(1)$, has a normal distribution with mean $b + \theta(1-t)$ and variance $(1-t)$. For a two-tailed test at level α, conditional power is then represented by the area of the crosshatched region above 1.96 (from Proschan et al. 2006). Reprinted with permission of Springer Science+Business Media LLC

Under the alternative hypothesis, $\theta = \theta_A$. Substituting θ_A in (5.1) gives the conditional power under the original alternative.

Under the null hypothesis, $\theta = 0$. Therefore, the conditional power under the null becomes simply:

$$\mathrm{CP}(t) = 1 - \Phi[(z_{\alpha/2} - b)/(1-t)^{1/2}]. \tag{5.2}$$

Under the current trend,

$$\mathrm{CP}(t) = 1 - \Phi[(z_{\alpha/2} - b/t)/(1-t)^{1/2}]. \tag{5.3}$$

A criticism often leveled at conditional power is that these three projections – under the null, under the alternative, and under the current trend – do not exhaust the possible future trajectories the data may take. This criticism fails to account for the flexibility conditional power allows, for one is not restricted to these three drift parameters. Many data monitoring committees find it useful to look at a range of drift parameters. As shown in Fig. 5.2, the conditional power is sensitive both to t and to drift. Nor is one restricted to projecting a constant drift; if there is reason to believe the effect size is likely to change in time, one can adapt the formulas to account for such change.

Some protocols suggest a threshold probability below which a DMC is encouraged to recommend stopping the trial. Obviously, if the conditional power against reasonable alternatives fell below the Type I error rate, everyone would

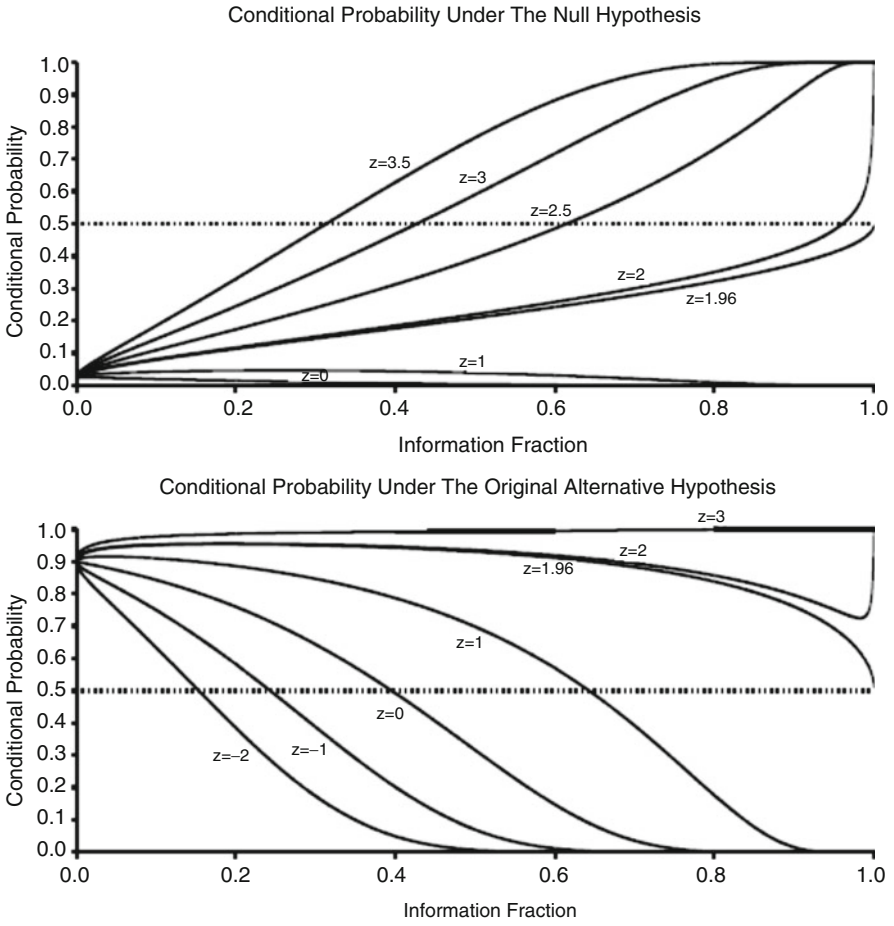

Fig. 5.2 Conditional power as a function of t and the z-score. Conditional power is not necessarily a monotonic function of information time t. For z-scores close to the critical value, conditional power is very sensitive to t (from Proschan et al. 2006). Reprinted with permission of Springer Science+Business Media LLC

consider the trial futile. We have seen protocols that recommend stopping the trial with conditional powers in the 40–50% range. In our opinion, these conditional powers are usually much too high; our own preference is to continue the trial unless the conditional power against reasonable alternatives is quite low (we prefer the 15–25% range). All this raises a question: what is a reasonable alternative? If the experimental condition is such that one can scientifically justify a belief in a constant drift parameter over the life of the trial (or proportional hazards if the trial is studying a time-to-event variable), then we suggest taking the optimistic end of the 80% observed confidence limit, projecting on the basis of that estimate, and recommending stopping only if the conditional power is below 20%. Pepe and Anderson (1992)

have also suggested using the optimistic end of a confidence interval to assess futility. A DMC contemplating recommending early stopping for futility should play out a host of reasonable alternatives consistent with the data and with prior hypotheses before recommending stopping for futility. Once a trial has ended, restarting it or beginning again is extremely difficult.

4.2 Predictive Power

Predictive power, a Bayesian alternative to conditional power, is essentially conditional power integrated over a distribution of future effect sizes. Rather than looking at a single drift parameter, Spiegelhalter et al. (1986) suggested averaging the conditional power over the distribution of the drift parameter. Before beginning the trial, the investigators specify a prior distribution for the treatment effect. An equivalent approach is to specify a prior distribution for the drift parameter θ. At the time of an interim analysis, the conditional power is averaged over the posterior distribution π of θ given the B-value. Predictive power accounts for the current data not only through CP(t) but also through the posterior distribution of θ (Freedman and Spiegelhalter 1989). Formally, the predictive power is:

$$\mathrm{PP}(t|b) = \int \mathrm{CP}(t,\theta)\pi\{\theta|B(t) = b\}\mathrm{d}\theta.$$

4.3 Group Sequential Designs

A group sequential method is another approach to assess futility. In group sequential designs, stopping boundaries can be calculated for efficacy (i.e., rejection of the null hypothesis $H_0 : \theta = 0$), as well as for futility (i.e., rejection of the alternative hypothesis $\mathrm{H_A} : \theta = \theta_\mathrm{A}$; Emerson and Fleming 1989). The error spending function methodology introduced by Lan and DeMets (1983), which consists of defining an α-spending function to terminate a study early if H_A is true, can be extended to define a β-spending function to terminate a study early if H_0 is true. Pampallona and Tsiatis (1994) have considered the general case in which an α-spending and a β-spending function are defined to ensure simultaneously that the Type I error does not exceed the prespecified significance level α and that the Type II error does not exceed the prespecified β (i.e., the power of the trial remains above 1-β). In such an approach, different stopping boundaries (or spending functions) can be adopted for efficacy and futility. It is often desirable to choose very conservative efficacy stopping boundaries (so that the trial is stopped early for efficacy only in the presence of extreme treatment benefits), but such conservatism may not always be warranted in the choice of futility stopping boundaries. In cases where investigators and sponsors can agree on a clinically unimportant effect size, the Emerson and Fleming (1989) method can be adapted to exclude that size.

4.4 Phase 3 Trial with Futility Determined By a Phase 2 Endpoint

Some Phase 3 trials embed a Phase 2 endpoint as a stopping rule for futility. In oncology, for example, if the objective response rate or some other early sign of activity falls below a predefined threshold (e.g., complete or partial response rate less than 10%), the likelihood that the treatment has enough activity to improve long-term clinical outcomes such as progression-free survival or overall survival may be too low to warrant continuation of the trial. An example from ophthalmology comes from the wet form of age-related macular degeneration where the primary outcome of many Phase 3 trials is visual acuity at 12 months. If few patients exhibit regression of leakage at 3 months, many investigators would question the value of continuing such a trial.

The above two examples use surrogate, or Phase 2, data to terminate a Phase 3 trial that is not likely to be successful for a primary efficacy endpoint that would require longer patient follow-up. This approach has the advantage of allowing a recommendation about terminating for futility earlier than is required for recommending termination for efficacy. Some researchers have suggested embedding a Phase 2 trial into a Phase 3 trial so that the transition between the two phases is *operationally seamless* – as opposed to performing a randomized Phase 2 trial followed by a separate Phase 3 trial. The simplest version of this approach consists of using a classical Phase 2 design to screen for activity based on response, and to calculate the sample size required for the final analysis based on the final outcome of interest. Since the purpose of the Phase 2 trial is only to stop for futility on the basis of lack of activity, no adjustment in the overall α level is required. One-stage or two-stage Phase 2 designs may be used (Fleming 1982; Simon 1989; Bryant and Day 1995; Sargent et al. 2001), as well as a "selection" design if several experimental arms are simultaneously screened (Simon et al. 1985). In all cases, the Phase 2 and the Phase 3 portions of the trial are designed independently of each other. Inoue et al. (2002) present a Bayesian method for expanding a Phase 2 trial to Phase 3 in oncology.

A different approach that is particularly useful in selecting one or more doses of a new investigational agent is to use an *inferentially seamless* design in which several doses are tested in the Phase 2 portion of the trial, and then to select only the most promising ones to continue in the Phase 3 portion (Bretz et al. 2009). Various designs have been proposed for this purpose, with or without adaptations of some design aspects at the end of the Phase 2 (Thall et al. 1988; Inoue et al. 2002; Jennison and Turnbull 2006). All these designs control of the overall significance level of the trial. Their main purpose is to stop some (but not all) doses, rather than to stop the trial for futility. For a concise review, see Jennison and Turnbull (2007).

4.5 Which Method to Use

The choice of a statistical futility method for a particular trial should depend on the characteristics of the trial, the concerns of regulators and Institutional Review Boards/ethics committees, and the comfort levels of sponsors and investigators. In general, conditional power is most useful when the parties involved agree on the extrapolation curve or when the choice of curve is deemed to make little difference in the outcome of the analysis. Predictive power is most useful when reliable prior knowledge from previous trials can be specified. Examples might include overall survival for breast cancer and colorectal cancer trials and titer response in anti-infective trials. A Bayesian approach to trial design and early stopping may also be quite effective in pediatric indications for which trials in adults have already been performed and provide a reasonable prior distribution for the anticipated effect of treatment. This approach is sometimes called "bayesian borrowing" (Malec 2001). We do not recommend predictive power for antidepressants, antihistamines, or antiepileptic drugs where considerable variation is seen from trial to trial (although true Bayesians might argue that the solution is to use a very diffuse prior). Group sequential methods might be used where there is little justification for conditional or predictive power or little agreement on how to implement these methods. Such might be the case for emerging interventions for which there is little experience, for example, such as targeted therapies in oncology that operate on novel pathways. Group sequential methods would also be used when investigators and sponsors feel it desirable for the interim futility analysis to be methodologically similar to interim efficacy analysis and when they agree on meaningful values of the efficacy parameter that are clinically unimportant. Reaching such agreement is a long-standing problem in the design of noninferiority trials where sample size is tied to this specification (D'Agostino et al. 2003). Using a Phase 2 endpoint for early stopping of a Phase 3 trial has considerable appeal partly because it requires fewer assumptions than other methods. Although seamless methods have considerable potential there is, at present, too little experience to render an evidence-based judgment of their utility.

5 Practical Considerations for Futility Analyses

5.1 Lag in Treatment Effect and Minimum Follow-Up

Several clinical issues occur in practice that may render formulaic interim analyses for futility misleading. To begin with, many treatments have a lag in effect. This lag may give the false impression that the efficacy of the experimental treatment is similar to that of the control at the time of interim analysis. The VA-HIT trial (Robins et al. 2001) provides a good example of such a lag. The trial compared gemfibrozil to placebo with respect to the occurrence of fatal and nonfatal myocardial infarction.

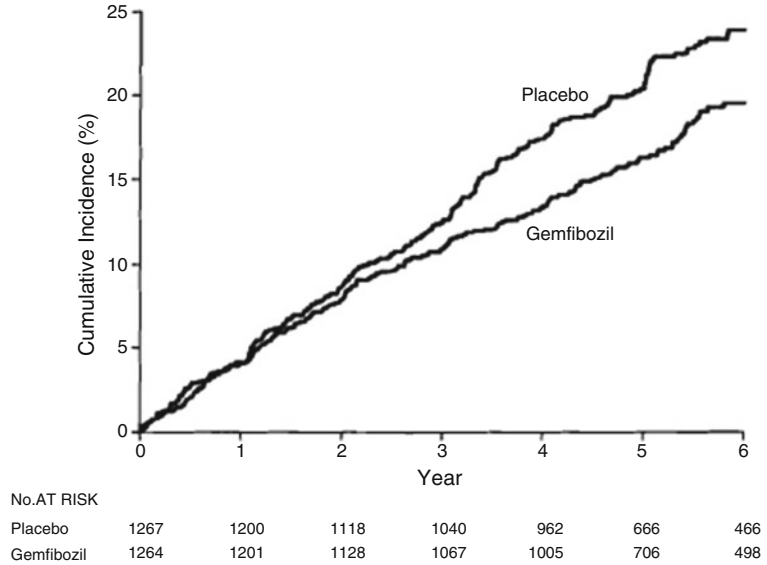

Fig. 5.3 Incidence of death from coronary heart disease and nonfatal myocardial infarction in the gemfibrozil and placebo groups (from Rubins et al. 1999). Reprinted with permission of the Massachusetts Medical Society

Participants in the trial were men with normal LDL-cholesterol levels but low levels of HDL-cholesterol. The hypothesis underlying the trial was that the HDL-raising effect of gemfibrozil would lead to a decreased risk of nonfatal myocardial infarction or death due to coronary heart disease. Lipid-altering drugs, however, may take months or even years to change the coronary arteries in a manner sufficient to lead to a meaningful reduction in observed risk. The designers of the trial built a lag to the effect of gemfibrozil into their sample size calculation. Thus, after 2 years, the event rates in the gemfibrozil and placebo groups were nearly identical, the Data Safety Monitoring Board did not recommend stopping. By the end of the trial, the curves had separated noticeably (22% risk reduction, $P = 0.006$, Fig. 5.3).

For other indications demonstrating that efficacy observed early can be maintained over time may be important. In these cases, futility analyses may be scheduled even after an intervention provides some evidence of efficacy and safety. Trials studying treatment of venous ulcers, for example, may show evidence of complete epithelization of the wound, but the durability of the healing is the ultimate outcome for assessing efficacy (Falanga and Sabolinski 1999). In such cases, efficacy would be defined as "heal and hold" (i.e., complete epithelization maintained for, say, 3 months). Futility analysis would be applied to the "heal and hold" outcome rather than the "heal" outcome.

5.2 Time Patterns

In some trials, patients recruited early differ from patients recruited later. Very often the sicker patients enter early and the less sick, or more recently diagnosed, enter later. Such difference can also occur if a protocol amendment changes eligibility requirements during the trial.

In trials that use a complex regimen or that employ a medical device, the investigators have a learning curve that may affect efficacy directly or indirectly. An example of a direct effect is learning the best way to instruct psychiatric patients on using a vagus nerve stimulator (Rush et al. 2005). An indirect effect may be through developing strategies to reduce toxicity so that fewer doses are missed, such as reducing the incidence of endophthalmitis in patients treated for age-related macular degeneration (Gragoudas et al. 2004).

5.3 Unreliability of Early Data

Early in the trial, important prognostic factors may be unbalanced because the sample size is small, and therefore the balance of baseline factors between treatment arms is highly influenced by the play of chance. If the prognostic factors happen to favor the control group, a DMC may be reluctant to recommend stopping the trial for futility for fear of missing an effect that may manifest itself after the chance imbalances have resolved. In some cases, the DMC may also be concerned about the completeness and reliability of the data. In on-going trials, it is almost inevitably the case that the data the DMC reviews are neither clean nor complete. Hence, the DMC may feel that the data upon which futility would have to be declared are too unreliable to make a recommendation to stop early. We illustrate these issues in the case study of Sect. 5.8.

5.4 Different Outcomes for Futility and Efficacy

In practice, the approach to futility will depend on the study. In futility analyses, complications occur when a different outcome is preferred for futility than for efficacy. This inconsistency may not pose a problem if the preferred futility outcome fulfills most of the properties of a surrogate for the efficacy outcome (Burzykowski et al. 2005); however, few surrogate outcomes have been established, and therefore stopping for futility is often decided on the basis of an intermediate endpoint that may or may not be a good predictor of the clinical endpoint. In rheumatoid arthritis trials, the FDA usually requires the Ritchie score (Lewis et al. 1988) as the primary efficacy outcome. Investigators, however, may consider morning stiffness the most important clinical outcome for the product. Similarly, sponsors may judge morning systems as the most important outcome for marketing. Therefore, the investigators,

or the sponsor, might want to terminate for futility on the basis of morning stiffness even though the latter has not been formally established as a surrogate for the Ritchie score.

In trials that use an active control to study a new treatment for seasonal allergic rhinitis, the primary efficacy outcome might be a score computed from symptoms in a patient's diary but the desirable variable for futility might be occurrence of moderate or worse headache because the investigators, or sponsor, might judge that an experimental treatment that cannot show improvement over the active control in headache would not be clinically important or viable in the market regardless of its efficacy with regard to average symptom score. Here, the designers of the trial might want to schedule the futility analysis earlier than the interim analysis of efficacy so as to limit the expense of the trial for a product that lacks sufficient clinical interest or market potential.

5.5 *Delayed Endpoints*

Most cancer trials, especially in the adjuvant setting, cannot stop for futility before recruitment has ended because the clinical outcomes, disease-free survival (DFS) and overall survival (OS), take a long time to occur. In such cases, futility would only affect the need to follow the patients further, but would not result in savings in terms of sample size.

Generally speaking, when overall survival time is the primary efficacy endpoint, interim analyses are unlikely to lead to early stopping for futility even if the null hypothesis is true. Goldman et al. (2008) considered using intermediate outcomes in this case. They investigated trials with overall survival as the primary outcome but with DFS and a composite outcome (OS or DFS, whichever occurs first) used for futility analysis. In this situation they simulated clinical trials using data consistent with the results of actual Southwest Oncology Group trials in advanced pancreatic cancer, advanced head and neck cancer, local prostate cancer, and local breast cancer. For all diseases studied, the authors found considerable utility in using these alternative endpoints for futility in terms of probability of early termination under the null hypothesis and average sample size. Pepe et al. (2009) suggested futility analysis to terminate trials of potential biomarkers when accumulating data indicate that the marker has inadequate sensitivity, specificity, or both.

5.6 *Timing of Futility Analyses*

In Sect. 5.4, we considered futility analysis that uses a Phase 2 endpoint in a Phase 3 trial with a Phase 3 primary efficacy outcome. In those trials, we assumed that if the Phase 2 futility variable did not point to trial termination then the trial would continue to its complete sample size. In this section, we consider the case where a

protocol calls for interim analysis of both futility and superiority. These analyses need not occur at the same time even if available software sometimes makes the simplifying assumption that they do. The investigators or sponsor might want to test superiority before or after deciding on futility especially if different outcomes are used for each. The interim analysis for superiority seeks a difference so large that continuation of the trial would be unnecessary to reject the null hypothesis. Some people consider continuation in such cases unethical. In an oncology trial where survival time is the primary efficacy endpoint sponsors might prefer to decide about futility after enough deaths have occurred on each treatment arm to be reasonably certain that further data would not change the decision.

Another approach for oncology trials might use 12-month survival for futility analysis but overall survival time as the primary efficacy variable. Here, if the 12-month survival meets some of the criteria for being a surrogate for survival time, the futility analysis would be scheduled after enough patients had at least 12 months of follow-up but the interim analysis of efficacy would take place after the required number of deaths had occurred, which could be earlier or later than the futility analysis.

Similarly, in an osteoporosis trial, a futility analysis of bone mineral density could be performed before an interim analysis of fracture incidence. The latter will require large sample sizes and considerable follow-up time while analysis of bone mineral density could occur much earlier. An intervention that appears futile on the basis of bone mineral density, which may be considered a "quasi-surrogate" for fracture, could stop early saving patient exposure and money. Similarly, an early analysis of ejection fraction might be relevant in a trial for studying prevention of myocardial infarction in heart failure patients. In trials of drugs with major toxicities, such as some drugs used in HIV, an analysis of viral load prior to an interim analysis of development of AIDS could be considered.

5.7 Binding vs. Nonbinding Futility Analyses

Throughout this chapter, we have danced around the words "guideline" rather than "boundary" to refer to the predetermined futility criteria. If the criteria to stop for futility are truly binding, then the "upper" or "superiority" boundary can be nudged down to account for the possible sample paths that are no longer in the rejection region leading to a gain in power (see, e.g., Lan and Trost 1997; Mehta 2005). When a boundary is declared binding, then the trial must stop for futility if the boundary is crossed. In that case, we are truly talking about a "boundary," and neither the DMC nor the sponsor has the option to ignore the fact that the boundary has been crossed. We strongly discourage binding futility boundaries because a committee needs to act on the basis of its collective judgment guided by, not determined by, statistical boundaries. Even if one could, we consider constraining the decisions of a DMC inappropriate. Upon observing that a futility boundary has been reached a DMC may opt to recommend continuing the trial anyway because the members see value in

continuing or they may feel there are still issues, like messy data, that preclude a firm recommendation to terminate. Alternatively, the DMC may declare that the futility boundary has been reached but the investigators or sponsor may choose to continue the trial anyway. In these cases, the "boundary" is indeed a "guideline."

In many trials sponsored by industry, the DMC reports directly to the industry sponsor, not to the investigators. In such cases, considerations about continuing a trial may relate in large measure to regulatory requirements and to marketing potential. For example, if the trial is multinational and the DMC reports to the sponsor, not to the investigators, and if the sponsor feels optimistic about approval of the drug in countries that use criteria different from those in the United States, the sponsor may wish to continue the trial even if the futility boundary is crossed. In other cases, the sponsor may judge that the drug still has marketing potential because of a favorable side effect profile compared to competing drugs. The sponsor may want to continue the trial to collect further data on safety and efficacy within subgroups of patients to plan a new trial. Conflict may arise between the DMC and the sponsor if the DMC recommends terminating the trial because the futility boundary was crossed and the sponsor brings up reasons, such as those just enumerated, to continue. We recommend specifying nonbinding guidelines in the trial protocol, in which case the DMC can decide whether or not to recommend stopping the trial based on risk-benefit considerations and not on business criteria. In many trials, the DMC speaks only to the investigators, not directly to the sponsor. (We are reminded of the John Bossidy's 1910 toast, "And this is good old Boston; The home of the bean and the cod; Where the *Lowells* talk to the *Cabots*; And the *Cabots* talk only to God").

In such cases, the considerations about stopping for futility differ from the considerations discussed above. A DMC needs to understand to whom it reports and what criteria it should use in considering early termination for futility.

As the trial progresses, the investigators or sponsor may feel a different projection method is more appropriate than the one written into the Statistical Analysis Plan. The method may be changed as long as the investigators and sponsor have not been unmasked or told that the futility boundary has been crossed. Indeed, the revised method may yield higher, or lower, conditional power than the original method. Although this reason for continuing is statistical in nature, it does not lead to any more bias than any of the other methods.

If the boundary has been crossed but the DMC recommends not stopping, the DMC should not inform the investigators or sponsor that the boundary was crossed until the end of the trial. For this reason, the DMC must be aware of the investigators' and sponsor's attitude toward the product being studied so that this decision is deemed not to be diametrically opposite to the one the investigators, or sponsor, might have made had there been no DMC. In trials sponsored by a large pharmaceutical company where the DMC reports directly to the sponsor, rather than to the investigators, the DMC can usually reveal the futility situation to a high-level executive not associated with the day-to-day operations of the trial. In a small company that has no other product on the market and a limited or nonexistent pipeline, this decision may be more difficult and the DMC would need some idea of the sponsor's likely position.

Table 5.4 Results of two planned interim analyses and the final analysis – see text for a description of the trial

	Treatment arm	
	Placebo	Active
Failures for primary endpoint		
First interim analysis	2/28 (7%)	9/31 (29%)
Second interim analysis	8/57 (14%)	15/59 (25%)
Final analysis	20/90 (22%)	18/90 (20%)
Deaths		
First interim analysis	0/28 (0%)	6/31 (19%)
Second interim analysis	8/57 (14%)	9/59 (15%)
Final analysis	24/90 (27%)	12/90 (13%)

Table 5.5 Distribution of poor prognostic factors at two planned interim analyses and the final analysis – see text for a description of the trial

	Treatment arm	
	Placebo	Active
Factor of poor prognosis # 1		
First interim analysis	5/28 (18%)	12/31 (39%)
Second interim analysis	14/57 (25%)	20/59 (34%)
Final analysis	23/90 (26%)	28/90 (31%)
Factor of poor prognosis # 2		
First interim analysis	2/28 (7%)	10/31 (32%)
Second interim analysis	13/57 (23%)	15/59 (25%)
Final analysis	20/90 (22%)	21/90 (23%)

Some sponsors may insist on a binding futility rule because if this product has a high probability of a negative outcome, the sponsor may wish to divert resources to other products in the pipeline. In our experience, DMCs that report to the investigators rarely use binding futility rules.

5.8 A Case Study

We now use an actual case study to illustrate some of the problems DMCs may face when they review interim analyses for futility. We have disguised the disease and treatment of the case study and modified the data slightly to preserve anonymity, but the situation described here is otherwise real. The disease in question was life-threatening and the standard therapy was known to have limited efficacy. The trial randomized 180 patients in a 1:1 ratio to receive an active drug or placebo. All received standard therapy as well. Tables 5.4 and 5.5 show the observed numbers of events for the primary outcome (failure to control the disease or death) and all-cause mortality, as well as the distribution of the most important prognostic factors for this

disease, at the times of the two planned interim analyses and the final analysis. At the first planned interim analysis, the difference between the randomized groups crossed the predefined futility boundary for the primary endpoint analysis, and more deaths were reported in the active treatment group (no death in the control group vs. 6 in the treated group, unadjusted $P = 0.025$). The DMC carefully reviewed these results, but the small numbers of events (Table 5.4) and the imbalances in the factors known to have strong prognostic impact in this disease (Table 5.5) led them to recommend continuation of the trial. At the second planned interim analysis, the difference observed for the primary outcome again crossed the predefined futility boundary, but the number of deaths in the two groups was almost the same (8 vs. 9 deaths, unadjusted $P = 0.85$). The DMC were still unimpressed by the difference in failures, given that the numbers of deaths were now similar in the two groups. In view of the absence of effective treatment for this frequently fatal disease, they reasoned that missing a potentially active therapy would be the least desirable outcome of the trial, and they recommended again its continuation. At final analysis, there were fewer deaths in the active treatment group (24 deaths in the control group vs. 12 in the treated group, unadjusted $P = 0.025$) even though the study failed to meet its primary objective.

Although this case study is anecdotal, it does show the extent to which a DMC may have to (or wish to) overrule predefined guidelines, if these were not carefully reviewed and discussed before the trial began. In the present example, one could indeed wonder (with the benefit of hindsight) whether it was wise to plan a first futility analysis based on only one-third of the information in a relatively small trial. Chance at such an early stage plays too large a role for almost any futility boundary to lead to a convincing recommendation to stop the trial. This example also illustrates that the DMC and the sponsor may value a negative result quite differently. Finally, the case study shows the extent to which data quality and other happenstances in trial conduct (such as accidental imbalances) form an integral part of the DMC's assessment, even thought the statistical boundaries cannot explicitly account for them.

5.9 The Role of Data Monitoring Committees

DMCs play a vital role in interpreting a futility analysis. Thus far in this chapter we have used the word "recommend" when we have spoken of the actions of the DMC, for those running the trial, not the DMC, are responsible for taking action. Crucial to its role is the method the DMC uses to report its recommendations. With regard to reporting, we have personally been involved in three types of trials. As shown in Fig. 5.4, communication flow can be quite complicated. In some trials, the DMC reports directly to the investigators, or an executive committee of the investigators; these DMCs never report to the sponsor whether the sponsor is an industrial firm, a government agency, or a not-for-profit organization. In this type of trial, when

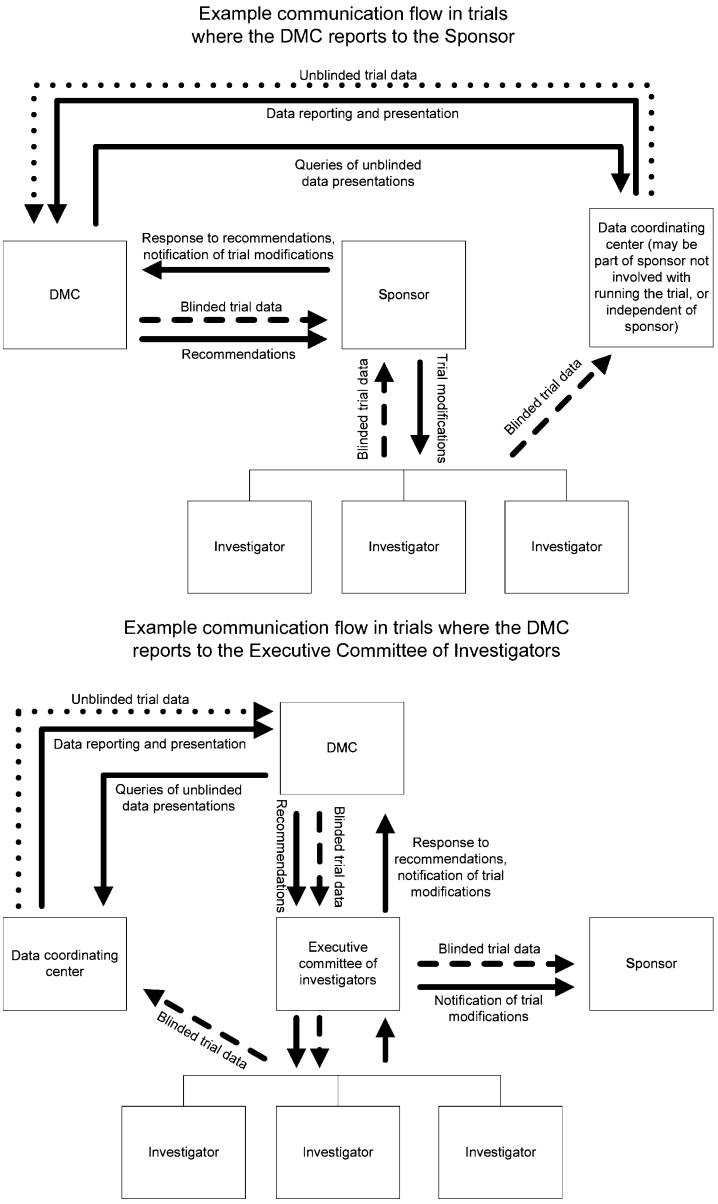

Fig. 5.4 Communication flow to report results of futility analysis

the DMC reports its recommendations, the investigators and the DMC discuss the recommendations and the investigators report their decision to the sponsor.

On the other extreme are trials in which the DMC reports directly to the sponsor and has no relationship to the clinical investigators. This type of trial is typical of

industry-sponsored trials aimed at regulatory approval of a product or an expansion of a label of an already marketed product. The discussion of whether, and how, to implement the DMC's recommendations occurs between the DMC and the sponsor.

In the third type of trial, a hybrid of these two, the DMC reports to a joint committee of investigators and sponsor, often called a Steering Committee. Although the DMC reporting lines may differ as shown in Fig. 5.4, the role of the DMC is essentially the same in all cases. It must ensure that the difficult decisions about continuing or stopping a trial are made consistent with sound science and in a manner totally independent of the self-interests of investigators and sponsor. The deliberations of the DMC should be regulated by a charter that clearly specifies the DMC's responsibilities for interpreting interim analyses, including futility analyses. The charter should specify if the futility analysis is binding or nonbinding. If the latter, some scenario planning would be desirable so that the DMC members and the decision-makers can be on the "same page" regarding relevant issues in deciding if termination for futility is necessary. The charter should indicate precisely whom to contact if the DMC recommends termination. If the DMC reports to the investigators, the charter should describe how the communication should occur. If the DMC reports to the industry sponsor, the charter should be unambiguous in describing who in the organization should be the contact. See Herson (2009) for a description of best practices for communications between sponsors and DMCs when the DMC reports directly to a sponsor.

6 Public Disclosure in Industry-Sponsored Trials

Public companies are required to disclose all material facts that may affect the value of the shares of their stock. Therefore, the law may require the sponsor to disclose the result of a futility analysis if the DMC reports it to the sponsor, regardless of the outcome and regardless of what action will be taken. Reaching a futility guideline and not stopping the trial can have disastrous effects on the valuation of a small public company. Continuing the trial after a public announcement of the futility criterion's being satisfied could raise the ire of patients and investigators. If the company is private, the group of investors on the board of directors would want management to report to them anything management knows about the trial. Thus, the outcome of a futility analysis could have a similar effect in a private company. For these reasons, we strongly recommend that if data cross the prespecified boundary but the DMC decides not to recommend termination, the DMC should simply tell the sponsor that it is recommending continuation of the trial. Only after the trial is over should the DMC inform the sponsor that a monitoring guideline had been reached. The DMC should be prepared to justify its reason for not having recommended stopping.

In the case of a publicly supported trial (e.g., the NIH in the US or the MRC in England), the sponsor need not inform the public at the time a trial is not stopped even though the futility boundary has been crossed (but see Sect. 5.7).

The fact that the DMC may make recommendations that are contrary to the prespecified guidelines is one reason why the sponsor (either industry or a government agency) must indemnify members of the DMC.

Closely related to public disclosure is reporting of futility analysis in the publication of trial results. A trial report and publication should indicate if the protocol included a planned futility analysis, if an ad hoc futility analysis was conducted instead of a planned futility analysis, what group conducted the analysis (most likely a data coordinating center), the statistical methods used (if any), the result of the futility analysis and what action (if any) was taken. This type of reporting is not routinely implemented. In a correspondence to a journal editor, Anderson (2007) claims that the paper by Hochman et al. (2006) reporting the results of a cardiovascular trial should have indicated that the trial was stopped for futility but, in reply, Hochman et al. (2007) claim that the trial's data monitoring committee did not in fact recommend terminating the trial for futility. This uncertainty is confusing to readers interpreting the data. Also, in reporting results if a trial is terminated for futility, the trial report should provide a complete enumeration of safety concerns. Trials terminated for futility may be considered "negative" but this should not preclude publication of results. The existence of the trial and its termination will be important information for researchers designing new trials and for those performing meta-analyses or writing systematic reviews.

7 Ethics

The investigators, the sponsor, and the DMC should read the informed consent document carefully to ensure that the participants in the trial understand that the trial may continue even when the data strongly suggest that the experimental treatment may not be shown to be more effective than the control arm.

Suppose the protocol specifies a futility boundary for efficacy but the investigators, the sponsor, or the DMC opts not to terminate the trial when the futility boundary is reached. An ethical problem appears if, at the end of the trial, the experimental treatment is found to be no better than placebo. The problem could be worse in a trial comparing an experimental treatment to an approved active control if the experimental treatment proves less effective than the active control. Futility should not be specified as the only termination criterion in active control trials unless investigators and sponsors are confident that the experimental treatment will not be shown to be inferior to active control.

In active control trials, the investigators and sponsor must decide before the trial begins whether the patients will be offered crossover treatment if a futility boundary is reached. Patients may want this flexibility if the treatments have different safety profiles, and the DMC may wish to consider the possibility of patients being crossed over in their assessment of futility. There is a growing literature on the ethics of early termination of clinical trials for futility as well as efficacy (Ashcroft (2001); Boyd (2001); Cannistra (2004); Evans and Pocock (2001); Iltis (2005); Lievre et al. (2001); Psaty and Rennie (2003); Trotta et al. (2008)).

8 Conclusions

The decision to include a planned interim futility analysis in a clinical trial protocol is complex. No "one size fits all" solution exists. We three authors object in principle to binding futility analyses. A protocol should include a futility analysis if there is a clear rationale for one – ethical issues, safety concerns are expected, rapidly developing competitive drugs, and an extensive pipeline of drugs for the same indication or if a futility result would have a profound effect on the practice of medicine. The previous sections have included some guidance for choosing a statistical method of futility analysis. In our experience, conditional power and group sequential methods work well. Predictive power can be used when Bayesian expertise and defendable prior information exists. Using a Phase 2 endpoint for early termination when this endpoint appears to signal poor efficacy results has not been used often but has considerable promise. In planning a trial investigators should certainly consider confounding issues such as lags in treatment effect, time patterns in data such as learning curves in treatment implementation, and differential quality in early compared to later data. The investigators, the sponsor, or the DMC may overrule futility decisions where confounding of this type is apparent.

Three examples illustrate the process of deciding on a futility analysis plan: a randomized clinical trial testing a drug for breast cancer, an anti-infective drug, and a mechanical heart valve implant. For a drug for breast cancer, futility analysis is useful if safety is a concern or a large pipeline of other candidate drugs is on the horizon. Conditional power would be a reasonable choice for a futility method. Assuming that overall survival is the primary efficacy endpoint for such a trial, investigators may want to use 12-month survival as the futility endpoint since, if overall survival were used for futility, with rapid accrual and long survival times, all patients may already be enrolled at the time of futility analysis. Anti-infective trials have short follow-up times (typically 14 days) and large sample sizes. A futility analysis early in the trial could result in considerable savings in cost and effort for the sponsor should the data point to futility. Here, conditional power, predictive power, or group sequential methods would probably all work well and yield similar results. Trials of a mechanical heart valve implant would be characterized by a short treatment phase and long follow-up. In a trial comparing a new valve design with an approved product, all patients are likely to be enrolled at the time of futility analysis. If the data raise no safety concern, investigators, the sponsor, and the DMC may want the trial go to completion even if a futility boundary were reached. In such cases, participating surgeons may have a learning curve in implanting the experimental device so futility analysis, if included at all, should not occur too soon. A group sequential approach to futility might be appropriate in this case if incidence of thromboembolic events were the primary outcome because regulators have a good idea of acceptable levels for such events (Grunkenmeier et al. 1994).

Decisions on recommending stopping for futility will often rest on the shoulders of the DMC. There is no substitute for adequate pretrial training of DMC members in the methods to be used, especially if these involve sophisticated

statistical tools as discussed in this chapter and other chapters in this book (Mehta 2010). In such cases, the sponsor is responsible for organizing pretrial meetings with the DMC to describe various scenarios that might occur during the trial and what actions might be taken. Herson (2009) indicates that it would be useful if organizations such as the Drug Information Association, Regulatory Affairs Professionals Association, Association of Clinical Research Professionals, British Institute for Regulatory Affairs, etc. provided DMC member training courses where potential DMC members could be certified for service on DMCs.

Acknowledgements We thank Catherine Conlon for preparing Fig. 5.4. We are grateful to David Sackett for his insightful comments on an earlier version of this chapter.

References

Abraham E., Reinhart K, Opal S et al (2003) Efficacy and safety of tifacogin (recombinant tissue factor pathway inhibitor) in severe sepsis. JAMA 290:238–247

Abraham E, Laterre, P-F, Garg R et al (2005) Drotrecogin-alfa (activated) for adults with severe sepsis and a low risk of death. New Engl J Med 353:1332–1341

Ajani JA, Winter KA, Gunderson LL et al (2008) Fluorouracil, mitomycin and radiotherapy vs. fluorouracil, cisplatin and radiotherapy for carcinoma of the anal canal: A randomized controlled trial. JAMA 299:1914–1921

Alling DW (1963) Early decision in the Wilcoxon two-sample test. J Am Stat Assoc 58:713–720

Alling DW (1966) Closed sequential tests for binomial probabilities. Biometrika 53:73–84; with correction note Biometrika 55:268 (1968)

Anderson JR (2007) Correspondence: Persistent coronary occlusion after myocardial infarction. New Engl J Med 356:1681

Ashcroft R (2001) Responsibilities of sponsors are limited in premature discontinuation of trials. BMJ 323:53

Boyd K (2001) Commentary: Early discontinuation violates Helsinki principles. BMJ 322:605–606

Bretz F, Koenig F, Brannath W et al (2009) Adaptive designs for confirmatory clinical trials. Stat Med 28:1181–1217

Bryant J, Day R (1995) Incorporating toxicity considerations into the design of two-stage Phase 2 clinical trials. Biometrics 51:1372–83

Burzykowski T, Molenberghs G, Buyse M (eds) (2005) The evaluation of surrogate endpoints. Springer, New York

Bussel JB, Cheng G, Saleh MN et al (2007) Eltrombopag for the treatment of chronic idiopathic thrombocytopenic purpura. New Engl J Med 357:2237–2247

Cairns JA, Wittes J, Wyse DG et al (2008) Monitoring the ACTIVE-W trial: Some issues in monitoring a noninferiority trial. Am Heart J 155:33–41

Cannistra SA (2004) The ethics of early stopping rules: Who is protecting whom? J Clin Oncol 22:1542–1545

Chuang-Stein C, Gallo PA et al (2006) Sample size reestimation: A review and recommendations. Drug Inf J 40:475–484

Curley MA, Hibberd PL, Fineman RN (2005) Effect of prone positioning on clinical outcomes in children with acute lung injury: A randomized controlled trial. JAMA 294:229–237

D'Agostino RB, Massaro JM, Sullivan LM (2003) Non-inferiority trials: Design concepts and issues – the encounters of academic consultants in statistics. Stat Med 22:169–186

Ellenberg SS, Fleming TR, DeMets DL (2002) Data monitoring committees in clinical trials: A practical perspective. Wiley, New York
Emerson SS, Fleming TR (1989) Symmetric group sequential designs. Biometrics 45:905–923
Emerson SS, Kittelson JM, Gillen DL (2007) Frequentist evaluation of group sequential clinical trial designs. Stat Med 26:5047–5080
Evans S, Pocock SJ (2001) Editorial: Societal responsibilities of clinical trial sponsors: lack of commercial pay off is not a legitimate reason for stopping a trial. BMJ 322:569–570
Falanga V, Sabolinski M (1999) A bilayered living skin construct (APLIGRAF) accelerates complete closure of hard-to-heal venous ulcers. Wound Repair Regen 7:201–207
Fleming TR (1982) One-sample multiple testing procedures for Phase 2 clinical trials. Biometrics 38:143–151
Freedman LS, Spiegelhalter DJ (1989) Comparison of Bayesian with group sequential methods for monitoring clinical trials. Contr Clin Trials 10:357–367
Gennari A, Amadori D, De Lena M et al (2006) Lack of benefit of maintenance paclitaxel in first-line chemotherapy in metastatic breast cancer. J Clin Oncol 24:3912–3918
Geyer CE, Forster J, Lindquist D et al (2006) Lapatinib plus capecitabine for HER2-positive advanced breast cancer. New Engl J Med 355:2733–2743
Goldman B, LeBlanc M, Crowley J (2008) Interim futility analysis with intermediate endpoints. Clin Trials 5:14–22
Gorelick PB, Richardson D, Kelly M et al (2003) Aspirin and ticlopidine for prevention of recurrent stroke in black patients: A randomized trial. JAMA 289:2947–2957
Gragoudas ES, Adamis AP, Cunningham ET et al (2004) Pegaptanib for neovascular age-related macular degeneration. New Engl J Med 351:2805–2816
Gray A, Raff AB, Chiriva-Internati M et al (2008) A paradigm shift in therapeutic vaccination of cancer patients: the need to apply therapeutic vaccination strategies in the preventive setting. Immunol Rev 222:316–327
Grunkenmeier GL, Johnson DM, Naftel DC (1994) Sample size requirements for evaluating heart valves with constant risk events. J Heart Valve Dis 3:53–58
Herson J (2009) Data and safety monitoring committees in clinical trials. Chapman and Hall, FL
Hochman JS, Lamas GA, Buller CE et al (2006) Coronary intervention for persistent occlusion after myocardial infarction. New Engl J Med 355:2395–2407
Hochman JS, Forman S, Reynolds HR (2007) Correspondence: Persistent coronary occlusion after myocardial infarction. New Engl J Med 356:1683
Iltis AS (2005) Stopping trials early for commercial reasons: The risk-benefit relationship as a moral compass. J Med Ethics 31:410–414
Inoue LYT, Thall PF, Berry DA (2002) Seamlessly expanding a randomized Phase 2 trial to Phase 3. Biometrics 58:823–831
International Conference on Harmonisation (1998). E9: Statistical Principles for Clinical Trials. http://www/ich.org/LOB/media/MEDIA485.pdf
Jennison C, Turnbull BW (2007) Adaptive seamless designs: selection and prospective testing of hypotheses. J Biopharm Stat 17:1135–1161
Jennison CM, Turnbull BW (2006) Confirmatory seamless Phase 2II/III clinical trials with hypotheses selection at interim: Opportunities and limitations. Biometr J 48:650–655
Kiel DP, Magaziner J, Zimmerman S (2007) Efficacy of a hip protector to prevent hip fracture in nursing home residents: the HIP PRO randomized controlled trial. JAMA 298:413–422
Lan KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. Biometrika 70:659–663
Lan K, Wittes J (1988) The B-value: A tool for monitoring data. Biometrics 44:579–585
Lan KK, Trost DC (1997) Estimation of parameters and sample size re-estimation. American Statistical Association Proceedings of the Biopharmaceutical Section
Lan KK, Simon R, Halperin M (1982) Stochastically curtailed tests in long-term clinical trials. Commun Statist Sequential Anal 1:207–219

Lanciano R, Calkins A, Bundy BN et al (2005) Randomized comparison of weekly cisplatin or protracted venous infusion of fluorouracil in combination with pelvic radiation in advanced cervix cancer: A gynecologic oncology group study. J Clin Oncol 23:8289–8295
Lee JJ, Feng L (2005) Review article: Randomized Phase 2 designs in cancer clinical trials: current status and future directions. J Clin Oncol 23:4450–4457
Levy MM (2004) PEEP in ARDS – How much is enough?. New Engl J Med 351:389–391
Lewis PA, O'Sullivan MM, Rumfield WR (1988) Significant changes in Ritchie scores. Rheumatology 27:32–26
Lievre M, Menard J, Bruckert E et al (2001) Premature discontinuation of clinical trials for reasons not related to efficacy, safety or feasibility. BMJ 322:603–606
Lorigan P, Verweij J, Papai Z et al (2007) PhaseIII trial of two investigation schedules of ifosfamide compared with standard dose doxorubicin in advanced or metastatic soft tissue sarcoma: A European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group Study. J Clin Oncol 25:3144–3150
Loughlin KR (2008) Immunotherapy: A new paradigm for androgen-refractory prostate cancer. Urol Oncol 26:575
Malec (2001) A closer look at combining data among a small number of binomial experiments. Stat Med 20:1811–1824
Mas J-L, Catellier G, Beyssen B et al (2006) Endaterectomy versus stenting in patients with symptomatic severe carotid stenosis. New Engl J Med 355:1660–1671
Mehta C (2005) EaST user manual, version 4. Cytel, Inc, Cambridge MA
Mehta C (2010) Sample size re-estimation for confirmatory trials. In:(Harrington D (ed) Current issues in the design of clinical trials. Springer, New York
Pampallona SM, Tsiatis AA (1994) Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. J Stat Plann Inference 42:19–35
Pepe MS, Anderson GL (1992) Two-stage experimental designs: Early stopping with a negative result. Appl Stat 41:180–190
Pepe MS, Feng Z, Longton G et al (2009) Conditional estimation of sensitivity and specificity from a phase 2 biomarker study allowing early termination for futility. Stat Med 28:762–779
Proschan MA, Lan KKG, Wittes JT (2006) Statistical monitoring of clinical trials: A unified approach. Springer, New York
Psaty BM, Rennie D (2003) Stopping medical research to save money: A broken pact with researchers and patients. JAMA 289:2128–2131
Robbins JA, Gensler G, Hind J et al (2008) Comparison of 2 interventions for liquid aspiration on pneumonia incidence: A randomized trial. Ann Internal Med 148:509–518
Robins SJ, Collins D, Wittes JT et al (2001) Relation of gemfibrozil treatment and lipid levels with major coronary events: VA-HIT, a randomized controlled trial. JAMA 285:1585–1591
Rubins HB et al (1999) Gemfibrozil for the secondary prevention of coronary heart disease in men with low levels of high-density lipoprotein cholesterol. New Engl J Med 341:410–418
Rush AJ, Marangell LB, Sackheim HA et al (2005) Vagus nerve stimulation for treatment-resistant depression: A randomized controlled acute phase trial. Biol Psychiatry 58:347–354
Sargent DJ, Chan V, Goldberg RM (2001) A three-outcome design for Phase 2 clinical trials. Contr Clin Trials 22:117–125
Simon R (1989) Optimal two-stage designs for Phase 2 clinical trials. Contr Clin Trials 10:1–10
Simon R, Wittes RE, Ellenberg SS (1985) Randomized Phase 2 clinical trials. Canc Treat Rep 69:1375–1381
Singh AK, Szczech L, Tang LT et al (2006) Correction of anemia with epoetin-alfa in chronic kidney disease. New Engl J Med 355:2085–2098
Spiegelhalter DJ, Freedman LS, Blackburn PR (1986) Monitoring clinical trials: Conditional or predictive power? Contr Clini Trials 7:8–17
Spriggs DR, Brady MF, Vaccarello L (2007) Phase 3 randomized trial of intravenous cisplatin plus a 24 or 96-hour infusion of paclitaxel in epithelial ovarian cancer: A Gynecologic Oncology Group study. J Clin Oncol 25:4466–4471

Stadler WM, Rosner G, Small E (2005) Successful implementation of the randomized discontinuation trial design: An application to the study of the putative antiangiogenic agend carboxyaminomidazole in renal cell carcinoma – CALGB 69901. J Clin Oncol 23:3726–3732

Sternberg CN, Petrylak DP, Sartor O et al (2009) Multinational double-blind phase III study of prednisone and either satraplatin or placebo in patients with castrate-refractory prostate cancer progressing after chemotherapy: The SPARC trial. J Clin Oncol 27(32):5431–5438

Teerlink JR, McMurray JJV, Bourge RC (2005) Tezosentan in patients with acute heart failure: design of the Value of Endothelin Receptor Inhibition with Tezosentan in Acute Heart Failure Study (VERITAS). Am Heart J 150:46–53

Thall PF, Simon R, Ellenberg SS (1988) Two-stage selection and testing designs for comparative clinical trials. Biometrika 75:303–310

Trotta F, Apolone G, Garattini S et al (2008) Stopping a trial early in oncology; for patients or for industry? Ann Oncol 9:1347–1353

US Food and Drug Administration (2006) Guidance for trial sponsors: Establishment and operation of clinical trial data monitoring committees. http://www.fda.gov/cber/gdlns/clintrialdmc.pdf

US National Heart, Lung and Blood Institute ARDS Clinical Trials Network (2004) Higher versus lower positive end-expiratory pressures in patients with the acute respiratory distress syndrome. New Eng J Med 351:327–336

Wittes J, Barrett-Connor E, Braunwald E et al (2007) Monitoring the randomized trials of the Women's Health Initiative: the experience of the Data and Safety Monitoring Board. Clin Trials 4:218–234

Chapter 6
Molecular Gene-Signatures and Cancer Clinical Trials

Mei-Ling Ting Lee

1 Introduction

Over the last 10 years, the process to develop molecular biomarkers and genomic tests for assessing the risk of cancer and cancer recurrence has been evolving. High-throughput technologies have increased the rate of discovery of potential new markers and facilitated the development of composite gene signatures that provide prognostic or predictive information about tumors. The traditional method to assess the risk of cancer recurrence is based on clinical/pathological criteria. The conventional design has been challenged, especially when the diseases may be heterogeneous due to underlying genomic characteristics.

Recently, there has been an increase in cancer clinical trials using gene signatures to assess cancer aggressiveness. For example, in some breast cancer studies, it was hypothesized that by using newly developed gene-signature tools one can identify subgroup of patients who will respond significantly to postsurgery (adjuvant) chemotherapy. Future treatments can then be designed to the individual person receiving it, and therefore spare the side effects of treatment to a large subgroup of potentially nonresponsive patients. On the other hand, a parallel goal is to identify what is the best treatment for patients: chemotherapy or hormonal therapy.

It is important to note that, if one of the major goals of using genomic biomarkers is to move closer to individualized treatment, the biomarkers or gene signatures need to be both prognostic and predictive.

Many studies with genomic biomarker and clinical investigations have been conducted in the past few years. In this article, we review these investigations and reproducibilities of the results.

M.-L.T. Lee (✉)
University of Maryland, College Park, MD 20742, USA
e-mail: mltlee@umd.edu

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7_6,

2 Phases, Guidelines, and Study Design in Development of Genomic Biomarkers

Pepe et al. (2001) defined phases for biomarker development for early detection of cancer. Phase 1 is for small preclinical exploratory studies and biomarker discovery; important questions to be asked in phase I are safety of treatment and/or pharmacokinetics for dosing. In Phase 2, a clinical assay based on specimens that can be obtained noninvasively is developed and evaluated for its clinical performance; Phase 2 asks whether there is any effect on surrogate measurements of the clinical outcome and whether the study should progress to Phase 3. Phase 3 is used more often in cancer to examine the sensitivity and specificity of a marker in predicting relapse. There are several markers now used this way, including BCR-Abl in CML, a number of blood chemistries used in myeloma, the gene signatures in breast cancer, etc. Phase 4 is to conduct prospective screening and to evaluate the sensitivity and specificity of the test on a prospective cohort. In Phase 4, a positive test may trigger a definitive diagnostic procedure which may be invasive and more expensive. A large cohort with long-term follow up is needed for Phase 4.

Skates and others (2003) discussed calculations of the risk of ovarian cancer from serial CA-125 values for preclinical detection in postmenopausal women. The CA-125 marker remains quite controversial. In cancer research, the validation and use of putative markers is a very difficult problem. Following the standard process of phase 1–4 may not be sufficient. Novel models that capture the association more precisely have a very large role.

Phase 5 evaluates the overall benefits and risks of the new diagnostic test on screened population. The phase structure has been adopted by the Early Detection Research Network (EDRN) and other research projects. Feng et al. (2004) discussed strategies and issues for genomic and proteomic biomarker discovery and validation. They recommended that biomarker development and evaluation should follow an orderly process and emphasized that although most biomarkers may not go through Phases 4/5 before their clinical adoption, these two phases are necessary so as to reduce the burden of health care cost and the potential harm to the affected public.

Brazma et al. (2001) published guidelines on minimum information about a microarray experiment (MIAME). This guideline has been widely adopted as a set of standards for microarray data. Simon (2005, 2007) developed guidelines for developing and validating therapeutically relevant genomic classifiers. Food and Drug Administration (2008) posted guidelines for industry on definitions for genomic biomarkers, pharmacogenomics, pharmacogenetics, genomic data and sample coding categories (E15); and a submission standard on genomic biomarkers related to drug response (E16). (See http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073162.pdfandhttp://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM174433.pdf).

Pepe et al. (2008) discussed standards for study design for evaluation of the accuracy of a biomarker used for classification or prediction. They propose a

prospective-specimen-collection, retrospective-blinded-evaluation (PRoBE) design in which biologic specimens are collected prospectively from a cohort that represents the target population that is envisioned for clinical application of the biomarker. Specimens and clinical data are collected in the absence of knowledge about patient outcome. After outcome status is ascertained, case patients with the outcome and control subjects without it are selected randomly from the cohort and their specimens are assayed for the biomarker in a fashion that is blinded to case-control status.

3 Emerging Molecular Gene-Signatures

After the previous update in 2000 (Bast et al. 2001), the American Society of Clinical Oncology (ASCO) published the 2007 update of recommendations for the use of tumor markers in breast cancer. Thirteen categories of breast tumor markers were considered in this 2007 guideline, six of which were new for the guideline. One of the new topics is multiparameter gene expression analysis for breast cancer. In this 2007 ASCO update, Harris et al. (2007) reported four assays, namely, the Oncotype DX, the MammaPrint test, the Rotterdam Signature, and the Breast Cancer Gene-expression Ratio. The other five new topics in the guidelines include urokinase plasminogen activator and plasminogen activator inhibitor1, cyclin E, proteomic analysis, bone marrow micrometastases, and circulating tumor cells as markers for breast cancer. The update also reported that some of the markers in the new categories do not have sufficient evidence to support routine use in clinical practice. Summarized below are descriptions of the four new multi-gene assays discussed in the ASCO 2007 report, as well as breast cancer gene expression ratio and the gene-expression grade index (GGI).

• **Strategies in the Development of Molecular Gene-Signatures**
Several genomic biomarkers and gene signatures have been investigated in the past few years. Dowsett and Dunbier (CCR Focus 2009) gave a thorough summary and review of traditional and emerging markers in personalized therapy for breast cancer. Among many methods that have been investigated, Sotiriou and Pusztai (2009) summarize three strategies for the development of a gene-expression prognostic signature in breast cancer: (1) top-down data-driven; (2) bottom-up hypothesis-driven; (3) candidate gene approach. They reported that the "top-down" data-driven approach was used in the development of MammaPrint (Agendia) where gene-expression data from cohorts of patients with known clinical outcomes are compared to identify genes that are associated with prognosis without any a priori biologic assumption. In the "bottom-up" hypothesis-driven approach, gene-expression patterns that are associated with a specific biologic phenotype or a deregulated molecular pathway are first identified and then subsequently correlated with the clinical outcome. The GGI also uses the bottom-up approach. The third strategy they considered is the "candidate-gene approach" where selected genes of

interest on the basis of existing biologic knowledge are combined into a multivariate predictive model. The Oncotype DX (Genomic Health) uses the candidate-gene approach for estimating outcome.

• **MammaPrint (70-Gene Signature)**
MammaPrint is a gene expression profile using a DNA microarray platform marketed by Agendia in the Netherlands. The test requires a sample of tissue that is composed of a minimum of 30% malignant cells. MammaPrint has received FDA clearance and is available in Europe and the United States.

The MammaPrint assay was developed on the basis of research conducted at the Netherlands Cancer Institute in Amsterdam and collaborating institutions (vant Veer et al. 2002; van deVijver et al. 2002). Using microarray technology and samples from lymph node-negative breast cancers, a dichotomous risk classifier was developed. Of 117 patients, 78 sporadic lymph-node-negative patients were selected to search for a prognostic signature in their gene expression profiles. Forty four Patients remained free of disease after their initial diagnosis for an interval of at least 5 years (good prognosis group), and 34 patients had developed distant metastases within 5 years (poor prognosis group). Using supervised classification method, approximately 5,000 genes were selected from the 25,000 genes on the microarray. The correlation coefficients of the expression for each gene with disease outcome were calculated. Two hundred and thirty one genes were found to be significantly correlated (correlation coefficient <-0.3, or >0.3) with disease outcome (distant metastases within 5 years). These 231 genes were ranked and the top 70 genes were selected. To initially validate the 70-gene profile, an additional set of tumors from patients free from distant metastases for at least 5 years after diagnosis and 12 tumors from patients with metastases within 5 years of diagnosis were analyzed. The 70-gene profile accurately predicted disease outcome in 17 of 19 patients, thereby confirming the initial performance of the prognostic classifier (Mook et al. 2007). A retrospective validation of the 70-gene profile was performed by the same research group using a consecutive series of 295 breast cancer patients (144 lymph node positive and 151 lymph node-negative, Cardoso et al. 2008).

Buyse et al. (2006) discussed an independent, albeit still retrospective, validation study of the MammaPrint 70-gene signature. Researchers used frozen archival tumor material from 302 node-negative patients from five non-Dutch cancer centers in three countries (the United Kingdom, Sweden, and France), who had similar characteristics to the Dutch initial series; all patients were 60 years or younger at diagnosis, were diagnosed before 1999, and had lymph nodenegative T1 or T2 breast cancer, and the great majority had not received adjuvant systemic therapy. The median follow-up of the TRANSBIG series was 13.6 years, which is much longer than the original series. Frozen samples were sent to Amsterdam for gene expression profiling using a custom-designed chip that assesses the mRNA expression of the 70-genes in triplicate (MammaPrint), and tumors were classified as high risk if the correlation coefficient for the average expression of the 70-gene profile was less than 0.4. Importantly, researchers in Amsterdam had no knowledge of the clinical outcome data. Central pathology review of ER status and grade was

carried out at the European Institute of Oncology (Milan, Italy) for nearly 80% of samples, and an independent audit was performed for all the clinical and pathologic data collected from the centers. Only an independent statistical office in Brussels, Belgium, had access to both the genomic and the clinico-pathologic data, and performed the validation analyses. The results of this validation study confirmed that the 70-gene profile was able to discriminate between patients at significantly high risk of distant metastases and death and patients with considerable low risk, and hence good prognosis, with hazard ratios of 2.79 (95% CI, 1.60–4.87) and 2.32 (95% CI, 1.35–4.0), respectively, for distant diseasefree survival (DDFS) and OS. The Mamma Print test was approved by FDA in 2007.

• **Oncotype DX (21-Gene Recurrence Score)**

Large clinical trials have demonstrated the benefit of tamoxifen and chemotherapy in women who have node-negative, estrogen-receptor positive breast cancer. However, since the likelihood of distant recurrence in patients treated with tamoxifen alone after surgery is about 15% at 10 years, at least 85% of patients would be overtreated with chemotherapy if it were offered to everyone. It has been shown that the likelihood of distant recurrence in patients with breast cancer who have no involved lymph nodes and estrogen-receptor-positive tumors is poorly defined by clinical and histopathological measures. Using the reverse-transcriptase-polymerase-chain-reaction (RT-PCR) technique, Paik et al. (2004) reported that the results of the assay of 21 prospectively selected genes in paraffin-embedded tumor tissue correlated with the likelihood of distant recurrence.

The Oncotype DX (Genomic Health) uses the 21 candidate-gene approach to estimate clinical outcome. The Oncotype DX was used for patients with node-negative, tamoxifen-treated breast cancer who were enrolled in the National Surgical Adjuvant Breast and Bowel Project clinical trial B-14 (entitled "A Clinical Trial to Assess Tamoxifen in Patients with Primary Breast Cancer and Negative Axillary Nodes Whose Tumors Are Positive for Estrogen Receptors"). The levels of expression of 16 cancer-related genes and 5 reference genes were used in a prospectively defined algorithm given by Paik et al. (2004) to calculate a recurrence score (RS) and to determine a risk group (low: RS$<$18; intermediate: 18$\leq$ RS $\leq$30: or high: RS $\geq$31) for each patient. This algorithm was established in a training set comprising mainly samples from the tamoxifen-alone arm of the NSABP-B20 trial and then validated in the tamoxifen arm of NSABP-B14. Genes associated with proliferation and endocrine responses are strongly represented. Although the recurrence score is a continuous measure of risk, it is conventionally used to identify three risk groups. Within the NSABP B-20 trial with a median follow-up of 10.9 years, the low-, intermediate-, and high-risk groups of tamoxifen-treated patients were associated with distant recurrence rates of $<$10%, 10–30%, and $>$30%, respectively. The recurrence score has been found to predict distant recurrence independent of age and tumor size, and is predictive of overall survival. With the use of the RT-PCR assay, Oncotype DX measures the expression of ER and HER2, as well as that of ER-regulated transcripts and several proliferation-related genes. Most of these genes are associated with outcome, and several can be assessed with the use

of conventional methods. Sparano and Paik (2008) reported that the Oncotype DX assay has been ordered for more than 40,000 patients and by approximately 6,000 different physicians since it became commercially available in January 2005.

• **Rotterdam 76-Gene Signature**
Using Affymetrix Human U133a GeneChips, Wang et al. (2005) analyzed frozen tumor samples from 286 lymph-node-negative patients who had not received adjuvant systemic treatment. They randomly divided the 286 samples (ER-positive and ER-negative combined) into a training set and a testing set. Based on a training set of 115 tumors, they identified a 76-gene signature consisting of 60 genes for patients positive for estrogen receptors (ER) and 16 genes for ER-negative patients. In an independent test set of 171 lymph-node-negative patients who had not received adjuvant systemic treatment, they found that this 76-gene signature showed 93% (52/56) sensitivity and 48% (55/115) specificity with AUC of 0.694 for the corresponding ROC curve. They claimed that this 76-gene signature provides a powerful tool for identification of patients at high risk of distant recurrence, and hence is useful for individual risk assessment in patients with lymph-node-negative breast cancer. The gene profile was informative in identifying patients who developed distant metastases within 5 years (hazard ratio 5.67 with 95% CI [2.59–12.4]), even when corrected for traditional prognostic factors in multivariate analysis (5.55 [2.46–12.5]).

• **Breast Cancer Gene Expression Ratio**
The Breast Cancer Gene Expression Ratio test (AvariaDx Inc, Carlsbad, CA) is a quantitative RT-PCR-based assay that measures the ratio of the HOXB6 and IL17BR genes. It is marketed as a marker of recurrence risk in untreated ER-positive/node-negative patients. The ratio of HOXB6:IL17BR genes was first reported by Ma et al. (2004, 2006) as predicting poor outcome in ER-positive patients treated with tamoxifen. The genes were discovered using an oligonucleotide array based on frozen material (Agilent Technologies, Santa Clara, CA) and subsequently validated by quantitative RT-PCR in archived material from the same tumor specimens.

• **97-Gene Gene-Expression Grade Index (GGI)**
Perou et al. (2000) originally reported a cluster of genes that correlated with cellular proliferation rates and was noted to have considerable variation between subgroups. Performing a supervised analysis, Sotiriou et al. (2006) defined a gene-expression grade index (GGI) score based on 97 genes. These genes were mainly involved in cell cycle regulation and proliferation and were consistently differentially expressed between low- and high-grade breast carcinomas. The GGI is an algorithm derived to evaluate the degree of similarity between a tumor sample and histologic grade.

The genomic test of MapQuant DxtM (Ipsogen, Marseilles, France and new Haven, CT, USA) is based on the genomic grade index (GGI). It directly measures the expression of the 97 genes that best characterize high-grade vs. low-grade tumors. Its company website claims that this test can resolve these “grade 2” tumors into either “grade 1” or “grade 3” tumors in 80% of cases.

• **Other Gene Signatures**
In addition to the above mentioned gene signatures, there are some other commercially available gene signatures, including Theros Breast Cancer indexsM (Biotheranostics, San Diego, CA). Sotiriou and Pusztai (2009) reviewed some of these gene signatures in breast cancer.

4 Trial Designs for Validation of Predictive Biomarkers

Biomarkers associated with disease outcome are referred to as prognostic markers and biomarkers associated with drug outcome are referred to as predictive markers. Mandrekar and Sargent (2009a,b) discussed clinical trial designs for predictive biomarker validation. In many cases, because of time and cost, retrospective validation is done using data from previously well conducted randomized controlled trials. Amado et al. (2008) reported that wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. This is an example of a marker that has been successfully validated using data collected from previous randomized trials. As a result of these studies, the label for panitumumab monotherapy for colorectal cancer has been restricted to a subgroup of patients with wild-type KRAS in Europe. On the other hand, the gold standard for predictive marker validation is a prospective randomized clinical trial. Summarized below are three prospective validation categories of designs and innovative adaptive designs reported by Mandrekar and Sargent (2009a,b).

• **Targeted or Enrichment Designs**
The targeted (enrichment) design is based on the paradigm that not all patients will benefit from the study treatment, but rather that the benefit will be restricted to a subgroup of patients who express (or not) a specific molecular feature. All patients are screened for the presence or absence of a marker (or gene signature), and only those with (or without) certain molecular features are included in the trial. Romond et al. (2005) employed an enrichment design and enrolled only human epidermal growth factor receptor 2 (HER2)-positive patients. They demonstrated that trastuzumub combined with chemotherapy significantly improved disease-free survival among women with surgically removed HER2-positive beast cancer. Subsequent analyses from NSABP B-31 (Paik et al. 2007; Perez et al. 2007) have raised the possibility of a beneficial effect of trastuzumab in a more broadly defined patient population.

• **Unselected Designs**
In the unselected design, all patients meeting the eligibility criteria (which does not include the status of a biomarker characteristic) are entered into the trial. Mandrekar and Sargent (2009a,b) note that the ability to provide adequate tissue may be an eligibility criterion for these designs, but not the specific biomarker result. These designs can be broadly classified into sequential testing strategy designs, marker-based designs, or hybrid designs, which are differentiated from each other

by the protocol specified approach to the prespecified type I and type II error rates (influencing sample size), analysis plans (including a single hypothesis test, multiple tests, or sequential tests), and randomization schema.

• **Hybrid Design**
When there is compelling prior evidence demonstrating the efficacy of a certain treatment for a marker-defined subgroup, it would be unethical to randomly assign patients with that particular marker status to other treatment options. Thus, patients in the other marker-defined subgroups are assigned the standard-of-care treatments. All patients are screened for marker status in this design. This design is powered to detect differences in outcomes only in the marker-subgroup that is randomized to treatment choices based on the marker status. Three large phase 3 marker validation trials have been launched with a hybrid design: the ECOG 5202 trial, the TAILORx trial, and the MINDACT trial.

• **Adaptive Designs**
In addition to the above designs, some novel statistical designs have been proposed recently. Freidlin and Simon (2005) proposed an adaptive design for randomized clinical trials of targeted agents in settings where an assay or signature that identifies sensitive patients is not available at the outset of the study. The design combines prospective development of a gene expression-based classifier to select sensitive patients with a properly powered test for overall effect. Wang et al. (2007) proposed an adaptive accural design and outlined a strategy to adaptively modify accural to two predefined marker-defined subgroups based on an interim futility analysis. Jiang et al. (2007) considered a biomarker-adaptive threshold design that is similar to the sequential testing strategy designs. Using a Bayesian hierarchical framework and based on outcomes from the accumulated data in the trial, Zhou et al. (2008) introduced an outcome-based adaptive randomization design to randomly assign patients to treatments in an adaptive fashion based on biomarker status.

5 Recent Examples of Cancer Clinical Trials with Genomic Biomarkers

5.1 MINDACT Trial

The Breast International Group (BIG), created in 1996 by European oncologists, involves research participants based in European countries, Canada, Latin America, Asia and Australasia. As an expansion of BIG, TRANSBIG is an international network that was created to promote translational research and international collaboration. The TRANSBIG consortium includes institutions in 21 countries from Europe and Latin America who are preparing a randomized clinical trial for the use of DNA chips to help determine therapies for lymph node-negative breast cancer patients (Hampton 2004). The first TRANSBIG molecular-based adjuvant trial for node-negative breast cancer patients is called MIcroarray for Node negative Disease

may Avoid Chemo-Therapy (MINDACT). It has been coordinated by the European Organisation for Research and Treatment of Cancer (EORTC) (see http://www.eortc.be/services/unit/mindact/MINDACT_websiteii.asp)

The MINDACT trial (EORTC 10041/ BIG 3-04) is a Phase 3 clinical trial conducted under the BIG and TRANSBIG network. With accrual started in February 2007, the trial will accrue 6,000 early breast cancer patients, either node-negative or with 1–3 positive lymph nodes.

Cardoso et al. (2008) review the MINDACT trial that compares the 70-gene signature MammaPrint (Agendia) with traditional clinical-pathological methods for assessing the risk of breast cancer recurrence in women with early breast cancer.

The EORTC website lists the following statistical design for primary tests in the MINDACT trial. In the group of patients who have a low-risk gene prognosis signature and high-risk clinical-pathological criteria, and who were randomized to use the 70-gene risk and thus receive no chemotherapy, a null hypothesis of a 5-year DMFS of 92% will be tested. With 6,000 patients accrued overall, this group has an expected size of 672 patients. With an accrual of 3 years, and a total duration of 6 years (so 3–6 years follow up for each patient), a one-sided test at 97.5% confidence level has 80% power to reject this hypothesis if the true 5-year DMFS is 95%. In the setting of the first randomization several other tests, comparing overall efficacy and chemotherapy assignment probabilities between the two prognosis methods, as well as within specific subgroups of discordant prognosis, will be important in assessing the value of prognosis according to the 70-gene signature.

It is hypothesized that using the genomic test in addition to traditional methods will result in more accurate risk assessment such that in the future 10–20% of patients could safely avoid adjuvant chemotherapy and its potential side effects. The MINDACT trial is the first in which microarray technology is being evaluated in a large, prospective clinical trial as a prognostic test to direct clinical decisions (Hampton 2004). In November 2008, with the accrual of the first 800 patients, the MINDACT trial reached its first milestone of the "pilot phase." The preliminary results of this phase demonstrate that the trial is logistically feasible. With an average accrual rate of 120 patients per month, the trial is expected to finish recruitment by the end of 2012 (EORTC Report 2009–2010).

In the MINDACT trial, mandatory fresh tumor samples for microarray analysis are stored for biobank and proteomic analysis. Also a representative paraffin tissue block has to be sent for central histopathology review and for the production of tissue microarrays. Optional blood sample collection may be performed for genetics and proteomics and for future research projects. Results of this trial could serve as a model for studies of microarrays as prognostic tools for many diseases.

Node-negative breast cancer patients who are particularly at risk of cancer recurrence can voluntarily participate in the MINDACT study. Patients who will undergo chemotherapy are randomly selected; a procedure that is important for the validity of the study and which provides the framework for determining whether the new genetic signature reliably indicates the risk of breast cancer relapse. All patients, whether receiving chemotherapy or not, are to be closely monitored. In the future, it is expected that 20% of women will be spared strenuous chemotherapy without any compromise in outcome.

5.2 TAILORx Trial

Sparano and Paik (2008) reviewed the development of the 21-gene assay (Oncotype DX, Genomic Health) and its clinical validation in the TAILORx trial.

The TAILORx trial is a Phase III randomized study of adjuvant combination chemotherapy and hormonal therapy versus adjuvant hormonal therapy alone in women with previous resected axillary node-negative breast cancer with various levels of risk for recurrence. Activated in 2006 and coordinated by ECOG (NCT00310180), the multicenter Trial Assigning Individualized Options for Treatment (TAILORx) was among the first to test the feasibility of a prognostic tool in clinical application. Using RT-PCR, the 21-gene assay was developed specifically for patients with ER-positive breast cancer and has been shown to predict distant recurrence.

In this trial, doctors will use a test called the Oncotype DX Breast Cancer Assay, which measures the activity of a set of genes in breast tumor tissue, to determine which women will receive adjuvant chemotherapy in addition to hormone therapy. With a null hypothesis of no difference, this trial uses a noninferiority design to determine whether patients with a recurrence score between 11 and 25 derive benefit from adjuvant chemotherapy with a larger type I error (one-sided, 10%) and smaller type II error (5%) than usual. A decrease in the 5-year disease-free survival (DFS) rate from 90% with chemotherapy to 87.5% or lower on hormone therapy alone would be considered unacceptable. All patients in the TAILORx trial will provide blood samples for banking and future research.

As of October 10, 2010, a total of 6,906 patients were randomized. The accrual of the TAILORx trial was hence closed (personal communication from Robert Gray). For details about the TAYLORx trial, see the trial webpage at http://www.cancer.gov/clinicaltrials/ECOG-PACCT-1.

5.3 SWOG S8814-INT0100 Trial

Under the hypothesis that the 21-gene recurrence score assay (Oncotype DX by Genomic Health) is prognostic for women with node-negative, ER-positive breast cancer treated with Tamoxifen alone and that a high RS predicts a large, additional chemotherapy benefit, the Southwestern Oncology Group conducted a phase III trial for postmenopausal women with node-positive ER-positive breast cancer (Albain et al. 2007). The two primary objectives of this trial were to determine if the recurrence score (1) provides prognostic data for disease free survival in the Tamoxifen-alone control arm and (2) predicts a group that does not benefit from chemotherapy followed by Tamoxifen, despite positive nodes.

The Southwest Oncology Group (SWOG)-8814, INT-0100 study was a phase 3, open-label, parallel-group, randomized controlled trial. Enrolled patients were randomly assigned in a 2:3:3 ratio to one of three drug regimens: (1) tamoxifen

alone (20 mg per day orally) for 5 years; (2) six cycles of CAF (cyclophosphamide orally on days 1-14, doxorubicin intravenously on days 1 and 8, and fluorouracil intravenously on days 1 and 8) followed by tamoxifen (CAF-T); or (3) CAF with concurrent tamoxifen (CAFT). There were 367 specimens (40% of the 927 patients in the tamoxifen and CAF-T groups) with sufficient RNA for analysis. The recurrence score was prognostic in the tamoxifen-alone group ($p = 0.006$; hazard ratio [HR] 2.64, 95% CI 1.33–5.27, for a 50-point difference in recurrence score). Results from the SWOG S8814 trial show that the recurrence score is prognostic for tamoxifen-treated patients with positive nodes and predicts significant benefit of CAF in tumors with a high recurrence score. A low recurrence score identifies women who might not benefit from anthracycline-based chemotherapy, despite positive nodes (Albain 2010).

5.4 ECOG E5202 Trial

The Eastern cooperative oncology group (ECOG) launched the randomized phase 3 study E5202 in patients with stage II colon cancer. Based on two molecular markers, patients with stage II colon cancer at a high risk for recurrence after surgery are randomly assigned to one of two arms: 5-FU, Leucovorin and Oxaliplatin with (versus without) bevacizumab. Patients with low risk for recurrence after surgery will not receive any adjuvant therapy; they were subject to observation only. The trial was opened nationally in August 2005, a total of 3,610 patients were to be accrued for this study within 5.8 years (http://www.cancer.gov/clinicaltrials/ECOG-E5202). Mandrekar and Sargent (2009a,b) mentioned that the design of ECOG E5202 trial will not allow for a determination of the benefit of bevacizumab in the low-risk strata. However, if the outcomes in the absence of treatment are as favorable as predicted in that group, no postsurgical therapy would generally be recommended.

Following the announcement of the negative results for bevacizumab in the Roche AVANT trial, however (in addition to the earlier negative results in NSABP C-08 study), ECOG had to stop the randomization and treatment with bevacizumab in E5202. Accrual is currently suspended while they decide if it would be worth accruing more patients to address marker-driven endpoints. By November 2010, a total of 1,008 low-risk and 918 high-risk patients have been entered in E5202 (personal communication from Robert Gray).

6 Comparisons of Gene Signatures

On the basis of a single data set of 295 samples and applying different gene-expression models, Fan et al. (2006) conducted pairwise comparisons of five gene-expression models: intrinsic subtypes, 70-gene profile (van deVijver et al. 2002; vant Veer et al. 2002), wound response, recurrence score (Paik et al. 2004),

and the breast cancer gene expression ratio. Despite the absence of gene overlap, the different gene models yielded similar predictions. They found that, although these gene signatures had little overlap in gene identity, the results were highly concordant. Comparisons showed that their sample predictions agreed in 77% of patients with ER+ cancer and 81% of all patients. These analyses suggest that even though there was very little gene overlap (the 70-gene and recurrence-score profiles overlapped by only 1 gene: SCUBE2) and different algorithms were used, the outcome predictions for the majority of patients with breast cancer would be similar.

Using the published TRANSBIG independent validation series of node-negative untreated primary breast cancer patients, Haibe-Kains et al. (2008) compared three gene expression signatures, the 70-gene, the Rotterdam 76-gene, and the 97-gene GGI. They reported that the three evaluated signatures had similar capabilities of predicting distant metastasis free survival.

Dowsett and Dunbier (2008) examined five prominent gene expression signatures and three emerging single biomarkers. They reported that, despite the fact that the GGI is composed almost exclusively of proliferation genes, the GGI scores correlate with the well-known luminal A and B classifications and to the risk groups produced using the OncotypeDx assay, the MammaPrint assay, and the 76-gene Rotterdam signature. This overlap in groupings could perhaps be seen as even more surprising given that the Rotterdam signature does not contain any of the genes in either Oncotype DX or MammaPrint, and was developed using patients unselected for age, tumor size, grade, or estrogen-receptor/progesterone-receptor status, with development of metastatic disease within 5 years as a supervision criterion. However, given that proliferation-associated genes feature strongly in all of these signatures, this association may reflect the dominance of proliferation as a prognostic factor in breast cancer.

Wirapati et al. (2008) conducted a meta analysis involving breast cancer gene studies. Results exhibited similar prognostic proformance.

7 Statistical Considerations

7.1 Flaws to Be Avoided in Microarray Studies

Using MEDLINE search and hand screening, Dupuy and Simon (2007) conducted a review of microarray studies for clinical outcomes published through 2004. Ninety studies were identified and their descriptive characteristics were presented. The statistical analyses for 42 studies published in 2004 were examined in detail. They found that 21 of them contained at least one of the following three basic flaws: (1) in outcome-related gene finding, an inadequate control for multiple testing; (2) in class discovery, a spurious claim of correlation between clusters and clinical outcome, made after clustering samples using a selection of outcome-related

differentially expressed genes; (3) in supervised prediction, a biased estimation of the prediction accuracy through an incorrect cross-validation procedure. Based on the common mistakes they found from these studies, they gave thorough guidelines and a check list of "Do's and Don't" for statistical analysis and reporting for clinical microarray studies. (Ransohoff 2007; Ransohoff et al. 2008) also emphasized that cross-validation cannot replace the need for totally-independent assessment.

7.2 Statistical Hypotheses in Adaptive Designs

In addition to the designs reviewed above, some innovative statistical designs have been proposed that use adaptive methods. In conventional clinical trials, the overall treatment effect is to be considered in the entire randomized study population. Wang et al. (2005) and Simon and Wang (2006) investigated two aspects of clinical trials involving genomic biomarkers. Wang et al. (2007) investigated approaches to evaluation of treatment effect in randomized clinical trials with genomic subgroups. They use a genomic biomarker (either single gene or composite signature) to classify patients into two mutually exclusive subgroups: namely, patients in the genomic marker positive group ($g+$); and patients in the genomic marker negative group ($g-$). In this setting, the overall treatment effect Δ is defined as the average effect for the entire randomized study population derived by weighting the treatment effect in each genomic subgroup by its sample size fraction. The treatment effect in the genomic marker positive subgroup $g+$ is Δ_{g+}. The authors then consider the following hypotheses for the overall population:

- H_{0a}: $\Delta = 0$.
- H_{1a}: $\Delta > 0$.

Similarly, in the genomic marker positive subset group $g+$, they define the following hypotheses:

- H_{0g+}: $\Delta_{g+} = 0$
- H_{1g+}: $\Delta_{g+} > 0$

When the genomic subset is prospectively planned a priori, the authors consider the following composite hypotheses for the study:

- $H_0 = H_{0a} \bigcap H_{0g+}$: $\{\Delta = 0$ and $\Delta_{g+} = 0\}$
- H_1: $\{\Delta > 0$ or $\Delta_{g+} > 0\}$.

The composite null hypothesis here is the intersection of the previous two null hypotheses, namely, the hypothesis that neither all patients nor the $g+$ subset of patients benefit from the new treatment.

Using a bivariate normal model and other methods, the authors then compare the adaptive design method with the nonadaptive-design method. Empirical type I error probabilities and power comparisons are presented. They suggest that it is desirable

to test the prespecified genomic subset, regardless of the outcome of testing for all eligible patients. Instead of the conventional approach that H_{0g+} is tested only if H_{0a} is rejected, they suggest building a multiplicity adjustment rule such that H_{0a} is tested only if H_{0g+} is rejected.

They report that if H_{0a} is rejected, but H_{0g+} is not, the treatment effect is asserted for all patients studied and subgroup effects are reserved for checking the consistency of the effect. When only H_{0g+} is rejected, the treatment effect appears to be only in the genomic subset and the genomic biomarker may be predictive of therapeutic response. When both H_{0a} and H_{0g+} are rejected, the genomic biomarker is likely to have a prognostic therapeutic effect if there is no qualitative interaction between the therapy and the genomic biomarker. However, if a qualitative interaction exists and depending on the prevalence of the genomic biomarker, the genomic biomarker may be considered predictive of treatment effect.

The above adaptive design considered by Wang et al. (2007) assumes that the patients can be accurately classified into two groups. When the binary genomic marker gives imperfect classification of a diagnostic assay, Maitournam and Simon (2005) discussed the impact on the efficiency of targeted clinical trials in terms of number of screened patients to meet planned numbers of subjects in the $g+$ and $g-$ subgroups for randomization. They modeled assay specificity and whether a treatment effect is expected for the less responsive group of patients.

7.3 Predictive Analysis of Clinical Trials

A predictive biomarker is a biological measurement made before treatment to identify which patient is most likely or least likely to benefit from a particular treatment. Simon (2010a,b) proposed that the development of treatments with predictive biomarkers requires major changes in the standard paradigms for the design and analysis of clinical trials. He outlined a prediction based approach to the design and analysis of randomized clinical trials. Freidlin and Simon (2005) also considered an adaptive signature design. Their threshold designs are adaptive, not with regard to sample size or randomization ratio, but rather with regard to the subset in which the new treatment is evaluated relative to the control. In the adaptive threshold design, they assumed that a predictive biomarker score was prospectively defined in a randomized clinical trial comparing a new treatment to a control. The score is not used for restricting eligibility and no cut-point for the score is prospectively indicated. Freidlin et al. (2010) demonstrated that the power of this approach can be substantially increased by embedding the classifier development and validation process in a K-fold cross-validation. On the other hand, risk prediction procedures are valuable tools for disease prevention and management. Uno et al. (2007, 2011) proposed methods for evaluating prediction procedures.

7.4 Multiple Testing, Sample Size and Power Calculations

Both the TAILORx and MINDACT trials discussed above are prospective clinical trials, but many studies use genomic convenient samples. Wang et al. (2007) investigated statistical issues in evaluating pharmacogenomics-based clinical effect for confirmatory trials and discussed the issue of imbalance, confounding, bias, design efficacy loss, type I and type II errors that may occur in the evaluation of convenient samples. While many studies published promising results and more comparison/validation studies are under way, it is important to emphasize that the dataset that is to be used for the validation study should not be obtained from the dataset used to develop the gene sets. Also, it is important to evaluate gene signatures on their original platform and compute them with the original algorithms on an independent population. Other important basic issues involved in microarray studies, including multiple testings, false discovery rates, sample size and power calculations, and machine learning methods are summarized in Simon et al. (2003) and Lee (2004).

Acknowledgements This research is supported in part by a Fund made to the University of Maryland from the Friedman Family of Potomac, MD.

References

Albain K, Barlow W, Shak S, Hortobagyi G, Livingston R, Yeh I, Ravdin P, Yoshizawa C, Baehner F, Davidson N, Sledge G, Winer E, Hudis C, Ingle J, Perez E, Pritchard K, Shepherd L, Allred C, Osborne K, Hayes D (2007) Southwest oncology group and the breast cancer intergroup of NA, San Antonio, TX 30th Annual San Antonio Breast Cancer Symposium Abstract #10

Albain KS (2010) Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: A retrospective analysis of a randomised trial. Lancet Oncol 11:55–65

Amado RG, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman DJ, Juan T, Sikorski R, Suggs S, Radinsky R, Patterson SD, Chang DD (2008) Wildtype KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol 26:1626–1634

Bast RC Jr, Ravdin P, Hayes DF, Bates S, Fritsche H Jr, Jessup JM, Kemeny N, Locker GY, Mennel RG, Somerfield MR (2001) 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: Clinical practice guidelines of the American society of clinical oncology. J Clin Oncol 19:1865–1878

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29:365–371

Buyse M, Loi S, vant Veer L et al (2006) Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Canc Inst 98:1183–1192

Cardoso F, Vant Veer L, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ (2008) Clinical application of the 70-gene profile: The MINDACT trial. J Clin Oncol 26:729–735

Dowsett M, Dunbier AK (2008) Emerging biomarkers and new understanding of traditional markers in personalized therapy for breast cancer. Clin Canc Res 14:8019–8026

Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Canc Inst 99:147–157

Fan C, Oh DS, Wessels L et al (2006) Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 355:560–569

Feng Z, Prentice R, Srivastava S (2004) Research issues and strategies for genomic and proteomic biomarker discovery and validation: A statistical perspective. Pharmacogenomics 5(6):709–719

Freidlin B, Simon R (2005) Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Canc Res 11:7872–7878

Freidlin B, Jiang W, Simon R (2010) The cross-validated adaptive signature design for predictive analysis of clinical trials. Clin Canc Res 16:691–696

Haibe-Kains B, Desmedt C, Piette F, Buyse M, Cardoso F, Van't Veer L, Piccart M, Bontempi G, Sotiriou C (2008) Comparison of prognostic gene expression signatures for breast cancer. BMC Genom 9:394

Hampton, R (2004) Breast cancer gene chip study underway. JAMA 291(24):2927

Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast RC Jr (2007) American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 25(33):5287–5312

Jiang W, Freidlin B, Simon R (2007) Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Canc Inst 99(13):1036–1043

Lee M-LT (2004) Analysis of microarray gene expression data. Kluwer, Dordrecht (now Springer)

Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, Muir B, Mohapatra G, Salunga R, Tuggle JT, Tran Y, Tran D, Tassin A, Amon P, Wang W, Wang W, Enright E, Stecker K, Estepa-Sabal E, Smith B, Younger J, Balis U, Michaelson J, Bhan A, Habin K, Baer TM, Brugge J, Haber DA, Erlander MG, Sgroi DC (2004) A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Canc Cell 5(6):607–616

Ma XJ, Hilsenbeck SG, Wang W, Ding L, Sgroi DC, Bender RA, Osborne CK, Allred DC, Erlander MG (2006) The HOXB13:IL17BR expression index is a prognostic factor in early-stage breast cancer. J Clin Oncol 24(28):4611–4619

Maitournam A, Simon R (2005) On the efficiency of targeted clinical trials. Stat Med 24(3):329–339

Mandrekar SJ, Sargent DJ (2009a) Clinical trial designs for predictive biomarker validation: One size does not fit all. J Biopharm Stat 19(3):530–542

Mandrekar SJ, Sargent DJ (2009b) Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. J Clin Oncol 27(24):4027–4034

Mook S, vant Veer LJ, Rutgers E et al (2007) Individualization of therapy using MammaPrint: From development to the MINDACT trial. Cancer Genomics Proteomics 4:147–155

Naoi Y, Kishi K, Tanei T, Tsunashima R, Tominaga N, Baba Y, Kim SJ, Taguchi T, Tamaki Y, Noguchi S (2010) Development of 95-gene classifier as a powerful predictor of recurrences in node-negative and ER-positive breast cancer patients. Breast Canc Res Treat. doi:10.1007/s10549-010-1145-z

Paik S, Shak S, Tang G et al (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New Engl J Med 351:2817–2826

Paik S, Kim C, Jeong J, Geyer CE et al (2007) Benefit from adjuvant trastuzumab may not be confined to patients with IHC 3+ and/or FISH-positive tumors: Central testing results from NSABP B-31. J Clin Oncol 25:18s

Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y (2001) Phases of biomarker development for early detection of cancer. J Natl Canc Inst 93(14):1054–1061

Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD (2008) Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: Standards for study design. J Natl Canc Inst 100:1432–1438

Perez EA, Romond EH, Suman VJ et al (2007) Updated results of the combined analysis of NCCTG N9831 and NSABP B-31 adjuvant chemotherapy with/without trastuzumab in patients with HER2-positive breast cancer. J Clin Oncol 25:18s

Perou CM, Sorlie T, Eisen MB et al (2000) Molecular portraits of human breast tumours. Nature 406:747–752

Ransohoff DF (2007) How to improve reliability and efficiency of research about molecular markers: Roles of phases, guidelines, and study design. J Clin Epidemiol 60(12):1205–1219

Ransohoff DF, Martin C, Wiggins WS, Hitt BA, Keku TO, Galanko JA, Sandler RS (2008) Assessment of serum proteomics to detect large colon adenomas. Canc Epidemiol Biomarkers Prev 17(8):2188–2193

Romond EH, Perez EA, Bryant J et al (2005) Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. New Engl J Med 353:1673–1684

Simon R (2005) Development and validation of therapeutically relevant multi-gene biomarker classifiers. J Natl Canc Inst 97:866–867

Simon R (2007) Development and validation of biomarker classifiers for treatment selection. J Stat Plann Inference 138(2008):308–320

Simon R (2010a) Clinical trials for predictive medicine: New challenges and paradigms. Clin Trials 7:516–524

Simon R (2010b) Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Personalized Med 7:33–47

Simon R, Wang SJ (2006) Use of genomic signatures in therapeutics development. Pharmacogenomics J 6:1667–1673

Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y (2003) Design and analysis of DNA microarray investigations. Springer, New York

Skates SJ, Menon U, MacDonald N, Adam N. Rosenthal, Oram DH, Knapp RC, Jacobs IJ (2003) Calculation of the risk of ovarian cancer from serial CA-125 values for preclinical detection in postmenopausal women, J Clin Oncol 21(10):206s–210s

Sotiriou C, Pusztai L (2009) Gene-expression signatures in breast cancer. New Engl J Med 360:790–800

Sotiriou C, Wirapati P, Loi S, et al. (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst 98:262–72

Sparano JA, Paik S (2008) Development of the 21-gene assay and its application in clinical practice and clinical trials. J Clin Oncol 26(5):721–728

Uno H, Cai T, Tian L, Wei LJ (2007) Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc 102(478):527–537

Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ (2011) On the C-statistics for evaluating overall adequacy of risk prediction prodecures with censored survival data. Stat Med (in press)

vant Veer LJ, Dai H, van de Vijver MJ et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536

van deVijver MJ, He YD, vantVeer LJ et al (2002) A gene expression signature as a predictor of survival in breast cancer. New Engl J Med 347:1999–2009

Wang SJ, O'Neill RT, Hung HM (2007) Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharm Stat 6(3):227–244

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365(9460):671–679

Wirapati P et al (2008) Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Canc Res 10:R65

Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ (2008) Bayesian adaptive design for targeted therapy development in lung cancer – a step toward personalized medicine. Clin Trials 5(3):181–193

Chapter 7
Targeted Clinical Trials

Stephen L. George and Xiaofei Wang

1 Background and Introduction

Randomized clinical trials (RCTs) are the most reliable scientific tool for the identification of safe and effective therapies for cancer. Indeed, it can be argued that such trials are indispensible for this purpose. The design and analysis of these trials are generally aimed at differences in overall population parameters. The trials compare outcome measures such as response rates, progression-free survival (PFS), disease-free survival, and overall survival (OS) for patients assumed to be representative of the population under study, determined by the specified eligibility criteria on the trial. Although no assumption of homogeneity of the patients studied or of the relative effects of treatment is required, major conclusions and, for new agents, subsequent regulatory approvals are generally based on the overall population. For example, a conclusion might be that some new therapy is better than a standard therapy in some population with respect to PFS, if the observed hazard ratio is significantly less than unity (in favor of the new therapy), even if the "average" benefit on other scales (e.g., difference in the PFS medians) is quite small. This is a legitimate and reasonable approach, one that has over time yielded major advances in cancer treatments. However, if the eligibility criteria are sufficiently broad, a generally desirable feature of traditional RCTs, unrecognized heterogeneity may have a substantial effect on the power of the trial (Betensky et al. 2002; Zhang et al. 2006). Indeed, there may be reasons to suspect that the treatment under study might benefit only a subset of patients in the population and

S.L. George (✉)
Duke University, Duke Box 2717, Durham, NC 27710, USA
e-mail: georg001@mc.duke.edu

X. Wang
Duke University, Duke Box 2721, Durham, NC 27710, USA
e-mail: xiaofei.wang@duke.edu

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7_7,

that any observed difference is due solely or largely to the results in this subset. If this is true and if such patients can be reliably identified beforehand, major benefits could accrue by targeting the subset population rather than the broader one. Benefits include more efficient clinical trials, avoidance of unnecessary treatment of patients who will not benefit from treatment, faster drug development times, and so on. Indeed, the concept of "personalized" medicine is predicated on the existence of such exploitable patient differences. Of course, the reliable identification of the subset population, assuming such exists, is not easy, but is of critical importance (George 2008).

The purpose of this chapter is to explore issues in the design and analysis of targeted clinical trials, in which a prespecified subgroup of patients, often based on a putative predictive biomarker, is identified and targeted for special attention. One important class of targeted designs involves "targeted" therapy, in which a particular biologic pathway is targeted by the therapy. In this case, patients who, say, overexpress a particular marker may be more sensitive (or less sensitive) to the targeted therapy than those who do not. However, the concept of a targeted clinical trial is applicable more broadly than to clinical trials of targeted therapy. Other examples of targeted subgroups include patients who are identified as being at especially high (or low) risk of failure for the therapies under study, whether or not these are targeted therapies, and patients who are identified as sensitive (or resistant) to certain classes of standard agents. The relative efficiency and other comparative performance characteristics of various designs in these settings will be explored and illustrated with specific examples.

2 Prognostic and Predictive Biomarkers

It is common to make a distinction between "prognostic" markers and "predictive" markers (Conley and Taube 2004; Lee et al. 2009). A prognostic marker is one that can be used to separate patients with a better prognosis from those with a worse prognosis when given some standard treatment. A predictive marker, on the other hand, is one that can be used to identify subsets of patients for whom the effect of some new treatment, relative to a standard or control treatment, varies among the subsets. Such a biomarker can thus be used to guide therapy for patients and to design clinical trials more efficiently. Given the recent explosion in techniques for molecular analysis and the strong interest in personalized medicine, there is no shortage of claims that a given biomarker is predictive for the use of specific agents. Some examples of these markers and the agents or therapies for which they are claimed to be predictive are given in Table 7.1.

These include estrogen receptor (ER) positivity and tamoxifen in breast cancer (Jordan 2006), HER2 receptor overexpression for trastuzumab in breast cancer (Madarnas et al. 2008) and in advanced gastric cancer(Van Cutsem et al. 2009)

Table 7.1 Some putative predictive markers

Marker	Agent	Cancer type
Estrogen receptor	Tamoxifen	Breast
HER2	Trastuzumab	Breast, gastric
HER2	Lapatinib	Breast
EGFR	Gefitinib	NSCLC
Cox-2	Celecoxib	NSCLC
CD133	Sorafanib, erlotinib	NSCLC
K-RAS	Cetuximab, panitumumab	Colorectal
Androgen receptor	Dasatinib	Prostate

and for lapatinib in breast cancer (Madarnas et al. 2008; Finn et al. 2009), EGFR mutations for gefitinib in lung cancer (Mok et al. 2009), Cox-2 expression for celecoxib in advanced NSCLC (Edelman et al. 2008), CD133+ circulating hematopoietic progenitor cells for sorafenib plus erlotinib in NSCLC (Vroling et al. 2009), K-ras mutations for cetuximab (Karapetis et al. 2008) and panitumumab (Amado et al. 2008) in advanced colorectal cancer, and an androgen receptor signature for dasatinib in castration-resistant prostate cancer (Mendiratta et al. 2009). There are many other such examples and the list is growing rapidly. Not all of the claims made for a predictive biomarker are based on the same strength of evidence. Indeed, one of the difficulties in this field is the lack of consistent application of accepted principles of methodology in the design and analysis of such studies. Although it is beyond the scope of this chapter to discuss these methodology issues in detail, there has been considerable work in this area (Biomarkers Definitions Working Group 2001; Pepe et al. 2001; Simon 2005a,b; Ransohoff 2007; Wang et al. 2007; Pepe et al. 2008; Simon 2008a, b, 2009; Taube et al. 2009; Wang and Zhou 2010; Wang et al. 2009) and guidelines have been proposed for the levels of evidence to be used to assess the clinical value of a tumor marker (Hayes et al. 1996; Locker et al. 2006; Harris et al. 2007). These are widely used but also widely ignored. In addition, several journals have incorporated rules for reporting the results of marker studies, the REMARK criteria (McShane et al. 2005; Hayes et al. 2006).

One important aspect of the study of biomarkers is the appropriate assessment of the marker and, for binary markers constructed from continuous measurements, the categorization into marker positive and marker negative groups. For example, in the HER2 case noted above, the usual definition of HER2 positivity requires an immunohistochemistry (IHC) score of 3+ and positive fluorescence in situ hybridization (FISH) (Gown 2008). Alterations at the DNA (amplification) and protein (overexpression) level usually occur in concert, and both FISH and IHC can be accurate methods to assess these alterations. But, as with all such assays, the tests are not perfect and the presence of false positives or false negatives and the inevitable missing data issues have implications for clinical trial design efficiency as discussed more below.

Table 7.2 Hazard rates

	M−	M+	Pooled
Trt 0	$\lambda_0(t)$	$\lambda_0(t)e^{\beta_2}$	$\lambda_0^{(p)}(t)$
Trt 1	$\lambda_0(t)e^{\beta_1}$	$\lambda_0(t)e^{\beta_1+\beta_2+\beta_3}$	$\lambda_1^{(p)}(t)$
Hazard ratio	e^{β_1}	$e^{\beta_1+\beta_3}$	$\lambda_1^{(p)}(t)/\lambda_0^{(p)}(t)$

2.1 Statistical Models

To clarify the definitions of prognostic and predictive markers, consider a simple proportional hazards model for a time-to-event outcome such as OS or PFS with two treatments and a single binary biomarker:

$$\lambda(t) = \lambda_0(t)\exp\{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2\}, \tag{7.1}$$

where $\lambda(t)$ is the hazard function for an event, $\lambda_0(t)$ is the baseline hazard function, X_1 is the treatment indicator (0 for the control treatment, 1 for the experimental treatment), X_2 is the biomarker indicator (0 for marker negative, 1 for marker positive), and X_1X_2 is the interaction term. The hazard rates and resultant hazard ratios are given in Table 7.2.

Thus, β_1 is a measure of the treatment effect (the log of the treatment hazard ratio when $X_2 = 0$), β_2 is a measure of the prognostic effect of the marker (the log of the marker hazard ratio when $X_1 = 0$), and β_3 is a measure of the predictive effect of the marker (a measure of the statistical interaction between treatment and biomarker, an expression of the differential treatment effect of the marker). A marker is said to be prognostic if $\beta_2 \neq 0$ and predictive if $\beta_3 \neq 0$. The most obvious way to test the predictive ability of a biomarker is to test $H_0 : \beta_3 = 0$ using standard statistical procedures. Unfortunately, such a test is generally underpowered unless the interaction is quite strong (Peterson and George 1993). Other statistical procedures have been proposed to check on the predictive strength of a marker (Bonetti and Gelber 2000; Byar 1985; Royston and Sauerbrei 2004; Vittinghoff and Bauer 2006; Pepe 2005).

In general, a marker may be prognostic and predictive, prognostic but not predictive, predictive but not prognostic, or neither prognostic nor predictive as indicated in Table 7.3.

To illustrate these possibilities more concretely, consider model (1) above with $\lambda_0(t) = 0.115$, and time measured in months, corresponding to an exponential time to failure distribution with a median of approximately 6 months, similar to the expected median survival in patients with stage IV nonsmall cell lung cancer (NSCLC). Also, suppose that $\beta_1 = -0.40$, corresponding to a treatment effect (hazard ratio) of approximately 0.67 in marker negative patients. Figure 7.1 gives the four possibilities from Table 7.2 with $\beta_2 = 0.50$ and $\beta_3 = -0.60$ for the cases where these quantities are not zero. These choices correspond to a situation in which, in the control patients, the M+ patients have a worse prognosis than M- patients (since

Table 7.3 Prognostic and predictive biomarkers

β_2	β_3	Prognostic?	Predictive?
0	0	No	No
$\neq 0$	0	Yes	No
0	$\neq 0$	No	Yes
$\neq 0$	$\neq 0$	Yes	Yes

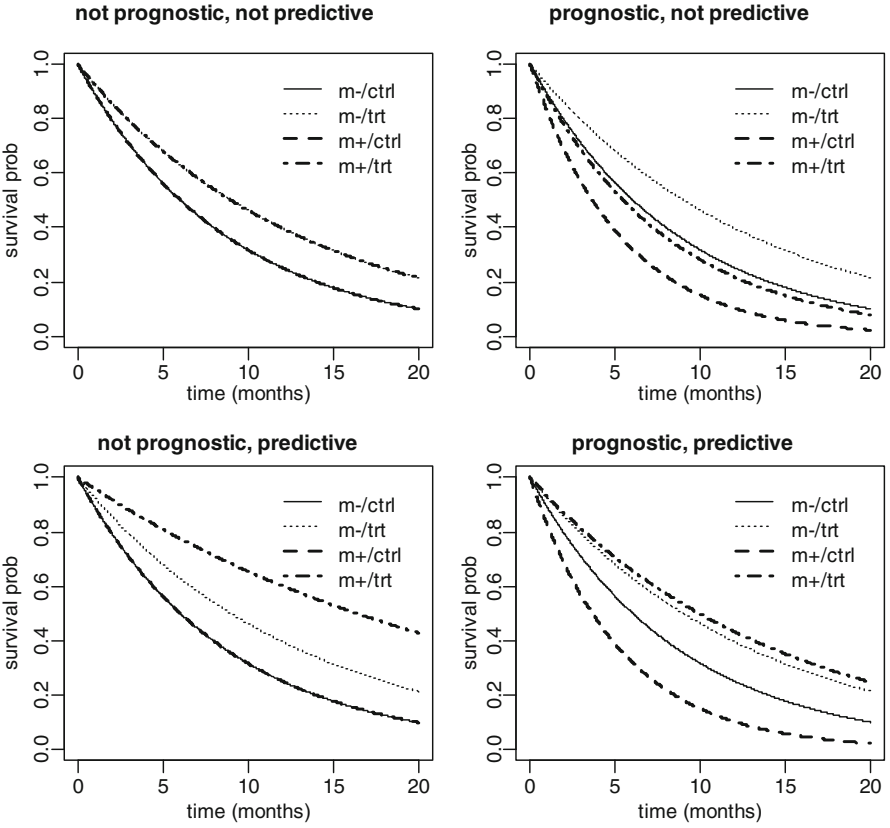

Fig. 7.1 Prognostic and predictive biomarkers

$\beta_2 > 0$) and in which the treatment advantage for the experimental treatment is larger in M+ patients than in M- patients (since $\beta_3 < 0$).

To be useful in designing targeted clinical trials, a marker must be predictive. It is neither sufficient nor necessary that it be prognostic. An important prognostic factor may be useful in setting eligibility criteria (e.g., by identifying particularly high-risk patients who *might* benefit from treatment) but, as seen above, the prognostic strength of a marker by itself is irrelevant to the issue of whether the treatment effect varies by marker subgroup. Also, even a predictive marker may not be very helpful if the predictive effect is not sufficiently strong. For example, if the treatment effect varies by biomarker category, but only slightly, this may or may not be helpful in

designing different regimens for the marker subgroups or in conducting targeted clinical trials. However, if the predictive value of the biomarker is high, particularly if the treatment effect is restricted exclusively to one of the biomarker categories, it is possible to gain substantial efficiency through the use of targeted designs. These concepts are made more precise in the next section.

3 Designs for Targeted Clinical Trials

It is intuitively obvious that if we have a reliably measured predictive biomarker, one that can readily distinguish patients likely to benefit from a therapy from those unlikely to benefit, we should be able to design much more efficient clinical trials. Issues related to this topic have been explored by many authors (George 2008; Simon 2008a,b; Simon and Maitournam 2004; Maitournam and Simon 2005; Jiang et al. 2007; Freidlin and Simon 2005; Hoering et al. 2008; Zhou et al. 2008; Mandrekar et al. 2005; Mandrekar and Sargent 2009a,b; Freidlin et al. 2008, 2010; Sargent and Allegra 2002; Sargent et al. 2005; Simon and Wang 2006). In this section, we discuss various designs that might be used in this setting and compare and contrast these designs with respect to hypotheses that can be tested, relative efficiency and related performance characteristics, and other issues. For simplicity, we assume, as above, that there are two treatments to be compared (control (C) and experimental (E)) in a randomized clinical trial and two biomarker groups (marker negative (M−) and marker positive (M+)) that are of interest. It is arbitrary how the marker groups are labeled and whether it is in the M+ or M− patients that the experimental treatment might be expected to do better, although in many examples, the M+ patients have both a worse prognosis and a larger treatment effect. More complex scenarios than two treatments and a single binary marker are of course possible, but the salient points can be made without the distractions of the added complexity.

3.1 Classes of Designs

There are four general types of designs that will be considered in this section:

- Traditional design – patients randomized without regard to the marker
- Targeted design – M+ patients randomized, M− patients not studied
- Biomarker stratified design – patients randomized within strata (M− and M+)
- Strategy design – patients randomized between two strategies (biomarker-directed, nonbiomarker-directed)

Variations of these basic design types are possible. For example, one might wish to expand a targeted design to include M− patients, treated with the control treatment, rather than simply excluding them from the trial. This would allow an evaluation of the prognostic, but not the predictive, effect of the biomarker. Also, in an

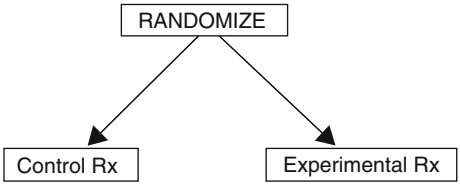

Fig. 7.2 Traditional RCT design

effort to increase efficiency, we will describe some "enrichment" strategies that are applicable to traditional designs and to biomarker-stratified designs.

3.2 Hypotheses Tested

The questions that can be addressed and hypotheses that can be tested with the various designs noted above are not identical. Relative efficiency issues are important but when two designs are not addressing the same questions and testing the same hypotheses, relative efficiency needs to be interpreted carefully. The types of hypotheses that can be addressed by the designs defined above are discussed in this section. Relative efficiency issues are discussed in the subsequent section.

3.2.1 Traditional Clinical Trial

Figure 7.2 gives the schema for a traditional clinical trial designed to compare an experimental treatment with a control treatment.

In a traditional clinical trial, patients are randomly assigned to the two treatment groups and the hypothesis tested is the equality of the treatments with respect to some primary outcome variable (often, OS or PFS in cancer clinical trials), typically without measurement of the biomarker at all. Of course, if the biomarker is not measured, obviously it is not possible to test prognostic or predictive hypotheses concerning the marker. However, if the biomarker is measured, such a trial differs from a biomarker-stratified design primarily in that the biomarker is not used as a stratification variable, although it may be used in the analysis. Thus, it would be possible to test hypotheses in a similar fashion to the biomarker-stratified design, with a potential loss of some efficiency because of the lack of stratification. The timing of measurement of the biomarker, assuming it is measured in a traditional design, is perhaps the major difference between a traditional and biomarker stratified design. In a traditional design, the biomarker may be measured at leisure without the time pressure required when it is required as a stratification variable. Indeed, as long as appropriate biological specimens are collected prior to the trial, retrospective assays, even those not planned initially, may be carried out. For such retrospective analyses, missing data may be a serious problem. Also, since the assays are

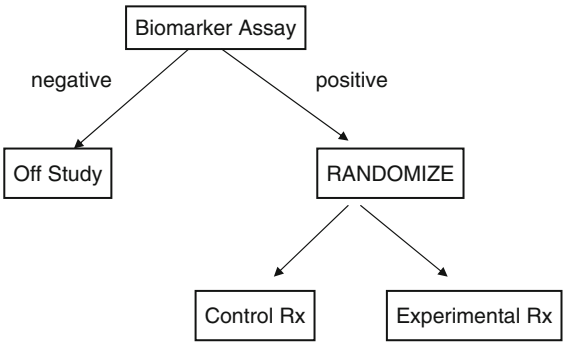

Fig. 7.3 Targeted RCT design

conducted after randomization, perhaps even after some of the primary outcomes are known, care must be taken to avoid introducing substantial bias.

3.2.2 Targeted Clinical Trial

Figure 7.3 gives the schema for a targeted clinical trial designed to compare an experimental treatment with a control treatment.

In a targeted clinical trial, patients are randomized between treatments only in the M+ group, since it is felt beforehand, presumably based on theoretical grounds and reliable empirical evidence, that these are the only patients who might benefit from the experimental treatment – or at least that there is likely to be very little benefit in the M− patients. In the basic targeted clinical trial formulation, M− patients are excluded from study, so the only possible hypothesis concerns the treatment effect *within* the M+ patients. As with a traditional design, when the marker is not measured, no hypothesis about the prognostic or predictive effect of the biomarker is possible. Such a trial would only be undertaken if one is reasonably confident about the lack of treatment effect in M− patients. In some settings, it may be reasonable to include M− patients, all treated with the control treatment. This would allow a test of hypothesis about the prognostic effect of the biomarker, but not the predictive effect, since no M− patients would receive the experimental treatment.

3.2.3 Biomarker Sratified Clinical Trial

Figure 7.4 gives the schema for a biomarker-stratified clinical trial designed to compare an experimental treatment with a control treatment.

In this design, the biomarker is used as a stratification variable, ensuring reasonable balance among treatments within the two biomarker groups (M+ and M−). Hypotheses about treatment effect as well as both the prognostic and predictive effect of the biomarker are possible. This design provides the best opportunity to

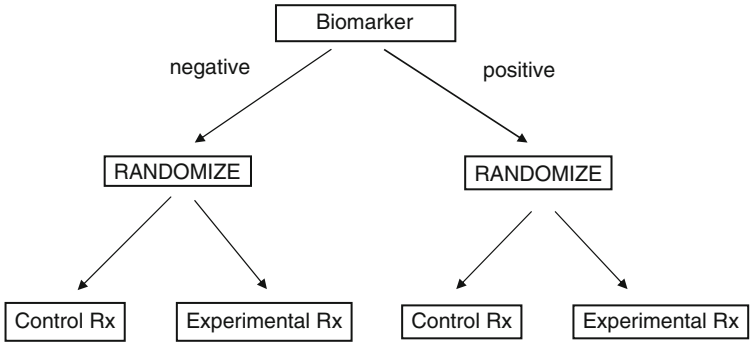

Fig. 7.4 Biomarker-stratified RCT design

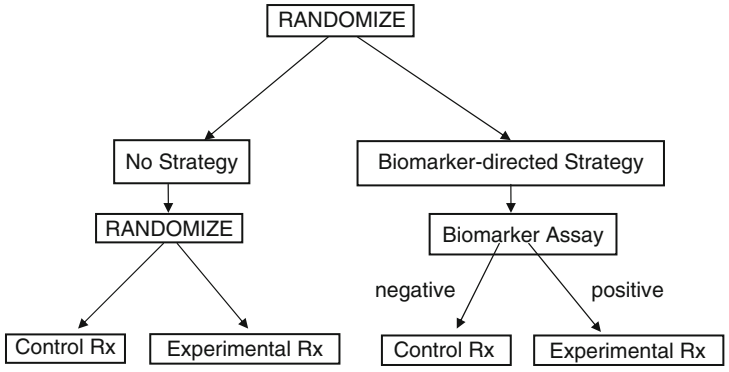

Fig. 7.5 Strategy design

test all of the relevant hypotheses about treatment effect and the prognostic and predictive effect of the biomarker.

3.2.4 Strategy Clinical Trial

Figure 7.5 gives the schema for a strategy clinical trial designed to compare a biomarker-based strategy approach with an alternative approach that does not use the biomarker to define treatment.

In a strategy design, patients are randomly assigned either to a *strategy* of using the biomarker in determining treatment assignment (a biomarker-directed strategy) or to a *strategy* of not using the biomarker in determining treatment (a nonbiomarker-directed strategy). The primary objective is to compare strategies, not treatments directly, although the treatment effect can be tested as a secondary objective in the no strategy arm if randomization is used as in Fig. 7.5. The primary interest is a pragmatic one of fundamental importance: Does the strategy of using

the biomarker information to determine treatment allow us to improve the outcome in a population of patients compared to not using the information? The details on the nature of the two strategies determine the specific hypotheses that can be tested. In one version of a strategy design, the one given in Fig. 7.5, the biomarker-directed strategy consists of giving the experimental treatment to M+ patients and the control treatment to M− patients, the nonbiomarker strategy randomizes patients to the two treatments in a $1-\gamma:\gamma$ ratio without regard to the biomarker value, where γ is the prevalence of M+ patients. The reason for randomization of the treatment assignment in the nonmarker-directed strategy, rather than treating all patients with the control therapy, is that if all patients were assigned to the control group, the biomarker-strategy group will do better if there is a treatment effect favoring the experimental therapy, regardless of whether the biomarker is predictive or not. But if the marker is not predictive, using a nonmarker-directed strategy of randomizing $100\gamma\,\%$ of the patients to the experimental therapy will work as well as the biomarker-strategy approach.

One difficulty with a biomarker-strategy design is that it requires a fairly large number of patients to address the primary objective of the trial. This is a direct result of the large overlap in treatment-biomarker subgroups within the two strategies being compared.

3.3 Relative Efficiency and Other Performance Characteristics

3.3.1 Sample Size and Number of Events

Even though the designs discussed above have varying objectives, it is of interest to investigate their relative efficiency. The most common measure of efficiency of a design is the required sample size, the number of patients required to meet the objectives (George 1984). For a clinical trial with full information on the primary endpoint available on each patient shortly after entry on study, the relative sample size required by two different designs is undoubtedly the most important measure of efficiency. However, in most phase III clinical trials in cancer, the primary endpoint is a time-to-event outcome such as PFS or OS. In this case, a more important consideration is the number of events, not the sample size. For example, the number of events required to compare two treatments with a 1:1 randomization and using a logrank test (two-sided, α-level test) in a proportional hazards setting is (George 2010)

$$\left(\frac{1}{d_1}+\frac{1}{d_2}\right)^{-1}=\left\lfloor\frac{(z_{\alpha/2}+z_\beta)^2}{(\ln\Delta_1)^2}\right\rfloor, \tag{7.2}$$

where $\lfloor x\rfloor$ denotes the smallest integer $\geq x, d_i$ is the number of events on treatment $i=1,2, z_x$ is the upper $100(1-x)$ percentile of the standard normal distribution, and $\Delta_1\neq 1$is the specified hazard ratio for which we desire the power of the test to be $1-\beta$. If $d_i\cong D/2$, where D is the total number of events, (7.2) can be rewritten as

$$D = \left\lfloor \frac{4(z_{\alpha/2} + z_\beta)^2}{(\ln \Delta_1)^2} \right\rfloor. \tag{7.3}$$

For a stratified design with two strata (M+ and M−), the required total number of patients N to be entered (Palta and Amini 1985) may be written as

$$N = \frac{4(z_{\alpha/2} + z_\beta)^2}{\{(\gamma V_1 \ln \Delta_1 + (1-\gamma) V_2 \ln \Delta_2)/\sqrt{\gamma V_1 + (1-\gamma) V_2}\}^2}, \tag{7.4}$$

where Δ_i is the hazard ratio in the ith stratum ($i = 1$ for M+ and $i = 2$ for M−), γ is the prevalence of M+ patients, and V_i is the probability of an event in the ith stratum. Thus, for the required number of events in a stratified design, the analogue to (7.3) is

$$D = \left\lfloor \frac{4(z_{\alpha/2} + z_\beta)^2}{(\gamma \ln \Delta_1 + (1-\gamma) \ln \Delta_2)^2} \right\rfloor. \tag{7.5}$$

Thus, the relative number of events required by the stratified design relative to the targeted design restricted to M+ patients is

$$\frac{D_s}{D_t} = \frac{(\ln \Delta_1)^2}{(\gamma \ln \Delta_1 + (1-\gamma) \ln \Delta_2)^2} = (\gamma + (1-\gamma)\theta)^{-2}, \tag{7.6}$$

where $\theta = \ln \Delta_2 / \ln \Delta_1$, or in terms of the model (7.1), $\theta = \beta_1/(\beta_1 + \beta_3)$. In the situations we are considering here, it is natural to suppose that $0 \le \theta \le 1$, where $\theta = 0$ corresponds to no treatment effect in the M− patients and $\theta = 1$ corresponds to an identical treatment effect in M+ and M− patients. Under the above assumptions, the required number of events in a stratified design is always greater than the number of required events in a targeted design (i.e., $D_s/D_t \ge 1$). The magnitude depends on the prevalence γ of M+ patients and the treatment effect θ in the M− patients relative to the M+ patients. For low prevalence and small $\theta, D_s/D_t$ can be quite large. If $\theta = 0$, the ratio is $D_s/D_t = \gamma^{-2}$, if $\theta = 1, D_s/D_t = 1$.

For illustration, we apply these results graphically to a specific clinical trial with a biomarker-stratified design, CALGB 30506, a phase III trial for stage I NSCLC patients in which 1,296 patients after surgical resection were to be randomized with equal probability, using a biomarker-stratified design, to adjuvant chemotherapy versus observation (Wang et al. 2009; Freidlin et al. 2010). Although the design was subsequently amended to drop the biomarker stratification, the original design and objectives of CALGB 30506 are used here since they are appropriate for our discussion. The randomization was stratified by risk group based on the model (High vs. Low) and pathological stage (IA vs. IB). Figure 7.6 gives the original design schema for this study.

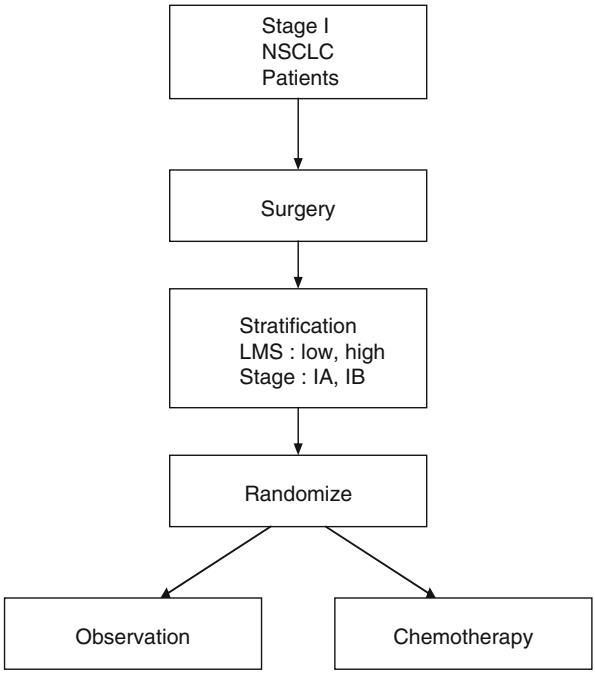

Fig. 7.6 CALGB 30506 design

The design assumptions were: equal allocation to chemo versus observation, an accrual rate of 288 randomized patients per year, the hazard rate (events per 100 patient years) for the low-risk patients on the observation arm is a constant $0.077, \theta = 0.45$, and $\gamma = 0.36$. In terms of model (7.1), these assumptions correspond approximately to $\lambda_0(t) = 0.077, \beta_1 = -0.203, \beta_2 = 0.728$, and $\beta_3 = -0.245$. Figure 7.7 gives plots of the $\log_{10}$ of the ratio of the required number of events for a stratified design to the required number of events for a targeted design as a function of θ for three different prevalence values: low $(\gamma = 0.10)$, medium (the design assumption, $\gamma = 0.36$), and high $(\gamma = 0.90)$. For the particular design assumptions on this trial $(\theta = 0.45, \gamma = 0.36)$, the ratio of required events is approximately 2.38.

The time required to achieve the required number of events for these two types of design is discussed in the next section.

3.3.2 Time to Completion of the Trial

Although the number of required events always favors a targeted design over a stratified design under the above assumptions, the lower accrual rate may require a longer accrual period and follow-up time to achieve the requisite number of events. Thus, the time to completion of the trial may be longer for a targeted design than for a stratified design. This "time to completion" metric is an important one in comparing designs in trials with a time-to-event outcome measure. It depends on the prevalence, prognosis, and relative treatment effect in the marker subgroups.

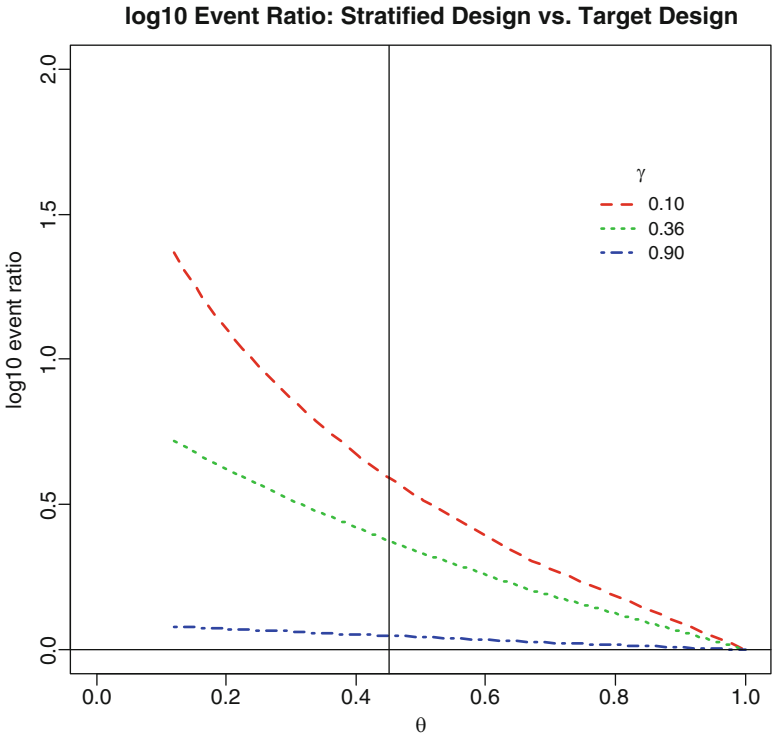

Fig. 7.7 Number of events for a stratified design relative to a targeted design

One problem with such a metric is that it is not uniquely defined by the required number of events. Thus, for a specified number of events, D, one could enter $N = D$ patients and follow all until failure or, since this might take a very long time, one could enter $N > D$ patients, and follow these patients until there are D events. In a targeted design we can write $N = D/V$, where V, the combined probability of an event (Schoenfeld 1981), depends on the accrual time necessary to accrue N patients (a function of the accrual rate) and the follow-up time necessary to achieve the required number of events (a function of treatment hazard rates and censoring rates). To illustrate how V is calculated in a specific situation, suppose that the failure distributions for the two treatments are exponential with hazard rates $\lambda_i (i = 1, 2)$, patients are accrued at a rate of a patients per year, the accrual period is T years, follow-up after the accrual ends is τ years, and there is no loss to follow-up. In this case, we can write V as (George and Desu 1974)

$$V = 1 - \frac{1}{2} \sum_{i=1}^{2} \frac{\exp(-\lambda_i \tau) - \exp(-\lambda_i (T + \tau))}{\lambda_i T}. \tag{7.7}$$

The planned length of study is thus $T + \tau$ years with the accrual and follow-up periods chosen so that D events are likely to occur during this time period. We can use any T and τ so long as T is large enough that at least D patients are entered and $T + \tau$ is long enough that there will be D events observed during this period. At one extreme, the minimum sample size required is equal to the required number of events, D, all patients followed to failure (assuming no loss to follow-up). That is, we let T be approximately D/a, so that the expected number of patients accrued will be D. But this strategy requires the longest follow-up time of any strategy to achieve D failures (George 2010; George and Desu 1974). The expected time until the last patient fails, a quantity related to the expected value of the nth order statistic from an exponential distribution (Xiangwei et al. 2003), is approximately $D/a + \log_e D/\lambda_1\Delta$, where $\Delta = \lambda_1/\lambda_2$, which is often so large as to be impractical. At the other extreme, one could enter patients continuously until the required number of events is obtained. This strategy produces the shortest time to achieve the requisite number of events (George and Desu 1974), but can lead to a very large sample size, particularly in situations with rapid accrual and low hazard rates, and with a high percentage of patients still in follow-up with no event at the time of the final analysis.

Obviously, some kind of compromise approach is needed between the two extremes discussed above. Although we desire a reasonably short time until the study is completed, it is also desirable to keep the excess number of patients entered over the required number of events to be relatively small. There is no unique way to do this. One approach is to fix the required follow-up time τ and solve for T so that the expected number of events (aTV) at $T + \tau$ is equal to the required number. The expected accrual (required sample size) using this approach is simply aT.

For a biomarker-stratified design with two strata (M+ and M−), the procedures are similar. The expected number of events for any given $T + \tau$ is $aT(\gamma V_1 + (1 - \gamma)V_2)$. To put the competing designs on a comparable basis, one could fix τ as above, and solve for the required T, following the steps described above for a targeted design, mutatis mutandis.

Figure 7.8 gives the relative total time to completion of the study, $T + \tau$, where $\tau = 3$ years, for a stratified design compared to a targeted design using the assumptions made for Fig. 7.7. As can be seen in Fig. 7.8, for moderate to large θ and for a small prevalence γ, this ratio can be less than 1.0. That is, the time to completion of a stratified design may be less than the time to completion of a targeted design, even though the required number of events is higher. For the specific assumptions used in CALGB 30506, $\theta = 0.45$ and $\gamma = 0.36$, this ratio is slightly above 1.0. Although not presented here, the results presented in Fig. 7.3 are not sensitive to the choice of τ, the curves are quite similar for choices of τ ranging from 0 to 10 years.

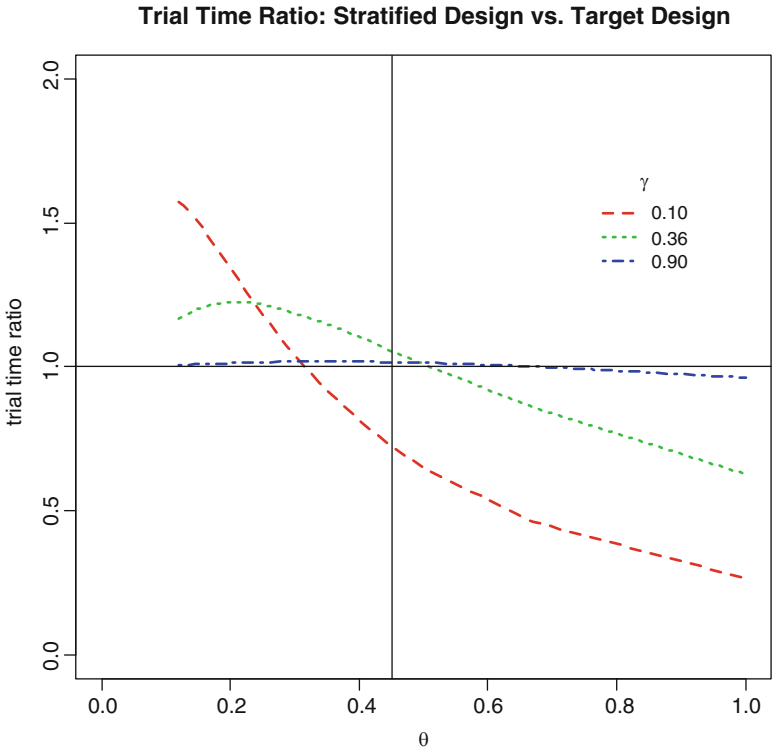

Fig. 7.8 Ratio of time required to achieve the required number of events for a stratified design relative to a targeted design

3.4 Other Considerations in Selecting Designs

In addition to the important considerations of hypotheses that can be tested and the relative efficiency considerations discussed above, there are other, more practical considerations that are important in devising an appropriate targeted clinical trial design. One of these is the complexity of the design. A clinical trial design that is simple to explain to potential participants and clear in its objectives is generally preferable to a more complex design, even if the more complex design has some advantages in scientific objectives or relative efficiency. The accrual rate may be adversely affected if the trial is difficult to explain. Thus, for example, the strategy design described above may present difficulties of this type.

Another practical difficulty in many of the biomarker-based designs considered here is that the marker must be assessed prior to registration on the trial. This obviously requires an assay that can be carried out quickly and leads potentially to a problem of missing data. If the assay sometimes leads to missing results, for whatever reasons, how are patients handled on the trial whose marker status is unknown – or whose marker status is unable to be determined quickly enough?

4 Examples of Designs

In this section, two recent clinical trials will be discussed to illustrate some of the previous concepts. The trials considered are in two areas: gefitinib (Iressa) in NSCLC and trastuzumab (Herceptin) in breast cancer.

4.1 Gefitinib (Iressa) Trials in NSCLC

The presence of activating mutations in the kinase domain of the epidermal growth factor receptor (EGFR) gene was first reported in 2004 (Yamamoto et al. 2009). In nonAsian patients with NSCLC, the percentage of patients with EGFR mutations is around 20%, but is much higher in female nonsmokers than in other patients. Asian patients have a higher percentage of mutations, around 50%, than other patients. Tumors exhibiting such mutations are highly sensitive to EGFR tyrosine kinase inhibitors. The first such TKI to be approved by the FDA was gefitinib (Iressa), which was approved in 2003 for patients with NSCLC under accelerated approval regulations. The approval was based on the basis of a small overall response rate, approximately 10%, in patients with advanced disease who were refractory to standard chemotherapy regimens. Follow-up studies failed to confirm any survival benefit for patients treated with gefitinib in unselected patient populations.

The mechanism of action of gefitinib suggests that targeting patients with EGFR mutations, or testing in a population with a highly enriched percentage of patients with such mutations, might be more productive than testing in an unselected population. The results of a recently reported trial, the Iressa Pan-Asia Study (IPASS), dramatically illustrate this point (Mok et al. 2009). IPASS was a traditional, phase III, multicenter, randomized, open-label, clinical trial comparing gefitinib with carboplatin plus paclitaxel as first-line treatment in clinically selected patients in East Asia who had advanced NSCLC. The primary hypothesis to be tested was that the gefitinib treatment was not inferior to the carboplatin–paclitaxel treatment with respect to the primary end-point, PFS. Testing for EGFR status prior to study entry was a not part of the study design, although patients were asked to provide consent for the use of their tumor samples in assessing EGFR status as part of subsequent planned exploratory analyses of EGFR and other biomarkers. Eligibility was restricted to East Asian patients with adenocarcinoma who were nonsmokers or former light smokers, restrictions certain to produce a relatively high percentage of patients with EGFR mutations (and a high percentage of females). Over 1,200 patients were enrolled, approximately 80% female and 60% with EGFR mutations, the latter percentage based on the 36% of patients who had evaluable EGFR data.

The overall results indicated that gefitinib was superior to carboplatin–paclitaxel with respect to PFS (hazard ratio $(\mathrm{HR}) = 0.74$). Further, in the subgroup of 261

patients who were positive for the EGFR mutation, the advantage for gefitinib was dramatic (HR = 0.48). Remarkably, in the subgroup of 176 patients who were negative for the mutation, the results were reversed: PFS was significantly longer among those who received carboplatin–paclitaxel (HR = 2.85). A formal treatment-EGFR interaction test was highly significant ($p < 0.001$).

4.2 Trastuzumab (Herceptin) Trials in Breast Cancer

The human epidermal growth factor receptor 2 (HER2) gene was identified in 1985 (Haq and Geyer 2009), and very soon thereafter amplification of this gene was demonstrated to be an important negative prognostic factor in breast cancer and to represent an important molecular subclass of breast cancer with a distinct clinical profile and response to systemic therapy (Bedard et al. 2009). Approximately 25% of patients with breast cancer cells overexpress this receptor or have a high gene copy number. These patients tend to relapse sooner and have shorter OS (Hudis 2007). Trastuzumab (Herceptin) is a monoclonal antibody directed against HER2 and a number of large clinical trials have shown a striking benefit for the use of trastuzumab as an adjuvant treatment for HER2-positive breast cancer patients (Madarnas et al. 2008; Slamon et al. 2001). In 2006, based on the combined results of two of these trials (Romond et al. 2005), NSABP B31 and NCCTG N9831, the FDA granted approval to trastuzumab as part of an adjuvant treatment regimen also containing doxorubicin, cyclophosphamide, and paclitaxel. Subsequent analyses have demonstrated that the benefit of trastuzumab is not limited to HER2 positive patients (Paik et al. 2008), emphasizing the importance of critically assessing the common assumption that the effect of a targeted agent is limited to patients who express the target.

5 Summary

Targeted clinical trials are likely to become increasingly more common in clinical cancer research. As emphasized in the topics covered in this chapter, these trials will need to be designed carefully to be certain that they address the correct hypotheses and are as efficient as possible. A biomarker-stratified design is likely the best choice when there is substantial uncertainty about the treatment effect in the marker subgroups. If there is substantial evidence that the treatment effect is limited primarily to a subset of patients defined by the marker, then a targeted design can be highly efficient. But there is no design that will be optimal in all cases and careful consideration of the issues described in this chapter will be required in specific cases.

Acknowledgements This work was supported in part by grants CA033601 and CA142538 from the National Cancer Institute.

References

Amado RG, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman DJ, Juan T, Sikorski R, Suggs S, Radinsky R, Patterson SD, Chang DD (2008) Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. J Clin Oncol 26:1626–1634

Bedard PL, de Azambuja E, Cardoso F (2009) Beyond trastuzumab: overcoming resistance to targeted HER-2 therapy in breast cancer. Current Canc Drug Targets 9:148–162

Betensky RA, Louis DN, Cairncross JG (2002) Influence of unrecognized molecular heterogeneity on randomized clinical trials. J Clin Oncol 20:2495–2499

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clin Pharmacol Therapeut 69:89–95

Bonetti M, Gelber RD (2000) A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data. Stat Med 19:2595–2609

Byar DP (1985) Assessing apparent treatment – covariate interactions in randomized clinical trials. Stat Med 4:255–263

Conley BA, Taube SE (2004) Prognostic and predictive markers in cancer. Dis Markers 20:35–43

Edelman MJ, Watson D, Wang X, Morrison C, Kratzke RA, Jewell S, Hodgson L, Mauer AM, Gajra A, Masters GA, Bedor M, Vokes EE, Green MJ (2008) Eicosanoid modulation in advanced lung cancer: cyclooxygenase-2 expression is a positive predictive factor for celecoxib + chemotherapy–Cancer and Leukemia Group B Trial 30203. J Clin Oncol 26:848–855

Finn RS, Press M, Dering J, Platek G, Arbushites M, Johnston S (2009) Progression-free survival (PFS) of patients with HER2-negative, estrogen-receptor (ER)-low metastatic breast cancer (MBC) with the addition of lapatinib to letrozole: Biomarker results of EGF30008. J Clin Oncol 27:15s.

Freidlin B, Simon R (2005) Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Canc Res 11:7872–7878

Freidlin B, Korn EL, Gray R, Martin A (2008) Multi-arm clinical trials of new agents: Some design considerations. Clin Canc Res 14:4368–4371

Freidlin B, McShane LM, Korn EL (2010) Randomized clinical trials with biomarkers: Design issues. J Natl Canc Inst 102:152–160

George SL (1984) The required size and length of a clinical trial. In: Buyse ME, Staquet MJ, Sylvester RJ (eds) Cancer clinical trials. Oxford University Press, Oxford, pp 287–310

George SL (2008) Statistical issues in translational cancer research. Clin Canc Res 14:5954–5958

George SL (2010) Design of phase III clinical trials. In: Kelly WK, Halabi S (eds) Oncology clinical trials: Successful design, conduct and analysis. Demos Medical Publishing, New York, pp 83–91

George SL, Desu MM (1974) Planning the size and duration of a clinical trial studying the time to some critical event. J Chronic Dis 27:15–24

Gown AM (2008) Current issues in ER and HER2 testing by IHC in breast cancer. Modern Pathol 21(Suppl 2):S8–S15

Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast Jr RC (2007) American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 25:5287–5312

Haq B, Geyer CE (2009) Role of trastuzumab in the adjuvant treatment of HER2-positive early breast cancer. Women's Health 5:135–147

Hayes DF, Bast RC, Desch CE, Fritsche H Jr, Kemeny NE, Jessup JM, Locker GY, Macdonald JS, Mennel RG, Norton L, Ravdin P, Taube S, Winn RJ (1996) Tumor marker utility grading system: A framework to evaluate clinical utility of tumor markers. JNCI J Natl Canc Inst 88:1456–1466

Hayes DF, Ethier S, Lippman ME (2006) New guidelines for reporting of tumor marker studies in breast cancer research and treatment: REMARK (letter). Breast Canc Res Treatment 100:237–238

Hoering A, LeBlanc M, Crowley JJ (2008) Randomized phase III clinical trial designs for targeted agents. Clin Canc Res 14:4358–4367

Hudis CA (2007) Trastuzumab – Mechanism of action and use in clinical practice. New Engl J Med 357:39–51

Jiang W, Freidlin B, Simon R (2007) Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. J Natl Canc Inst 99:1036–1043

Jordan VC (2006) Tamoxifen (ICI46,474) as a targeted therapy to treat and prevent breast cancer. Brit J Pharmacol 147(Suppl 1):S269–S276

Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, Simes RJ, Chalchal H, Shapiro JD, Robitaille S, Price TJ, Shepherd L, Au HJ, Langer C, Moore MJ, Zalcberg JR (2008) K-ras mutations and benefit from cetuximab in advanced colorectal cancer. New Engl J Med 359:1757–1765

Lee CK, Lord SJ, Coates AS, Simes RJ (2009) Molecular biomarkers to individualise treatment: Assessing the evidence. Med J Aust 190:631–636

Locker GY, Hamilton S, Harris J, Jessup JM, Kemeny N, Macdonald JS, Somerfield MR, Hayes DF, Bast Jr RC (2006) ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. J Clin Oncol 24:5313–5327

Madarnas Y, Trudeau M, Franek JA, McCready D, Pritchard KI, Messersmith H (2008) Adjuvant/neoadjuvant trastuzumab therapy in women with HER-2/neu-overexpressing breast cancer: a systematic review. Canc Treatment Rev 34:539–557

Maitournam A, Simon R (2005) On the efficiency of targeted clinical trials. Stat Med 24:329–339

Mandrekar SJ, Sargent DJ (2009a) Clinical trial designs for predictive biomarker validation: One size does not fit all. J Biopharmaceut Stat 19:530–542

Mandrekar SJ, Sargent DJ (2009b) Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. J Clin Oncol 27:4027–4034

Mandrekar SJ, Grothey A, Goetz MP, Sargent DJ (2005) Clinical trial designs for prospective validation of biomarkers. Am J Pharmacogenomics 5:317–325

McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, For the Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics (2005) Reporting recommendations for tumor marker prognostic studies (REMARK). JNCI J Natl Canc Inst 97:1180–1184

Mendiratta P, Mostaghel E, Guinney J, Tewari AK, Porrello A, Barry WT, Nelson PS, Febbo PG (2009) Genomic strategy for targeting therapy in castration-resistant prostate cancer. J Clin Oncol 27:2022–2029

Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y, Nishiwaki Y, Ohe Y, Yang JJ, Chewaskulyong B, Jiang H, Duffield EL, Watkins CL, Armour AA, Fukuoka M (2009) Gefitinib or Carboplatin-Paclitaxel in Pulmonary Adenocarcinoma. New Engl J Med 361:947–957

Paik S, Kim C, Wolmark N (2008) HER2 status and benefit from adjuvant trastuzumab in breast cancer (letter). New Engl J Med 358:1409–1411

Palta M, Amini SB (1985) Consideration of covariates and stratification in sample size determination for survival time studies. J Chronic Dis 38:801–809

Pepe MS (2005) Evaluating technologies for classification and prediction in medicine. Stat Med 24:3687–3696

Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y, (2001) Phases of biomarker development for early detection of cancer. J Natl Canc Inst 93:1054–1061

Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD (2008) Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Canc Inst 100:1432–1438

Peterson B, George SL (1993) Sample size requirements and length of study for testing interaction in a 2 x k factorial design when time-to-failure is the outcome. Contr Clin Trials 14:511–522

Ransohoff DF (2007) How to improve reliability and efficiency of research about molecular markers: Roles of phases, guidelines, and study design. J Clin Epidemiol 60:1205–1219

Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE Jr, Davidson NE, Tan-Chiu E, Martino S, Paik S, Kaufman PA, Swain SM, Pisansky TM, Fehrenbacher L, Kutteh LA, Vogel VG, Visscher DW, Yothers G, Jenkins RB, Brown AM, Dakhil SR, Mamounas EP, Lingle WL, Klein PM, Ingle JN, Wolmark N (2005) Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. New Engl J Med 353:1673–1684

Royston P, Sauerbrei W (2004) A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Stat Med 23:2509–2525

Sargent D, Allegra C (2002) Issues in clinical trial design for tumor marker studies. Seminars Oncol 29:222–230

Sargent DJ, Conley BA, Allegra C, Collette L (2005) Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 23:2020–2027

Schoenfeld D (1981) The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 68:316–319

Simon R (2005a) Development and validation of therapeutically relevant multi-gene biomarker classifiers. J Natl Canc Inst 97:866–867

Simon R (2005b) Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 23:7332–7341

Simon R (2008a) Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. Expert Opin Med Diagn 2:721–729

Simon R (2008b) The use of genomics in clinical trial design. Clin Canc Res 14:5984–5993

Simon R, Maitournam A (2004) Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Canc Res 10:6759–6763

Simon R, Wang SJ (2006) Use of genomic signatures in therapeutics development in oncology and other diseases. Pharmacogenomics J 6:166–173

Simon RM, Paik S, Hayes DF (2009) Use of archived specimens in evaluation of prognostic and predictive biomarkers. J Natl Canc Inst 101:1446–1452

Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J, Norton L (2001) Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. New Engl J Med 344:783–792

Taube SE, Clark GM, Dancey JE, McShane LM, Sigman CC, Gutman SI (2009) A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. JNCI J Natl Canc Inst 101:1453–1463

Van Cutsem E, Kang Y, Chung H, Shen L, Sawaki A, Lorente P (2009) Efficacy results from the ToGA trial: A phase III study of trastuzumab added to standard chemotherapy (CT) in first-line human epidermal growth factor receptor 2 (HER2)-positive advanced gastric cancer (GC). J Clin Oncol 27:18s

Vittinghoff E, Bauer DC (2006) Case-only analysis of treatment-covariate interactions in clinical trials. Biometrics 62:769–776

Vroling L, Lind JSW, de Haas RR, Verheul HMW, van Hinsbergh VWM, Broxterman HJ, Smit EF (2009) CD133+ circulating haematopoietic progenitor cells predict for response to sorafenib plus erlotinib in non-small cell lung cancer patients. Br J Canc 102:268–275

Wang X, Zhou H (2010) Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. Biometrics 66:502–511

Wang SJ, O'Neill RT, Hung HMJ (2007) Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. Pharmaceut Stat 6:227–244

Wang X, Pang H, Schwartz TA (2009) Building and validating high throughput lung cancer biomarkers. Chance 22:55–62

Xiangwei M, Jian G, Xinzheng W (2003) Order statistics for negative exponential distribution and its applications. IEEE International Conference Neural Networks and Signal Processing, pp 720–722

Yamamoto H, Toyooka S, Mitsudomi T (2009) Impact of EGFR mutation analysis in non-small cell lung cancer. Lung Canc 63:315–321

Zhang B, Li Y, Betensky RA (2006) Effects of unmeasured heterogeneity in the linear transformation model for censored data [Erratum appears in Lifetime Data Anal. 2007 Sep 13(3):431]. Lifetime Data Anal 12:191–203

Zhou X, Liu SY, Kim ES, Herbst RS, Lee JL (2008) Bayesian adaptive design for targeted therapy development in lung cancer – a step toward personalized medicine. Clin Trials 5:181–193

Chapter 8
Design Issues for Quality of Life Studies Subject to Dropout

Diane L. Fairclough

1 Introduction

Whenever a statistician is asked to participate in the design of a study in which dropout is anticipated either because of mortality/morbidity, side effects, or lack of efficacy, there are design issues that do not have simple solutions. The primary concern is that the missing data cannot be ignored if one wishes to obtain an unbiased estimate of the outcome. This chapter focuses on the questions that must be asked and options that the statistician has for the design of this kind of study when the outcome measure differs for subjects who continue on the study in contrast to those who dropout. Typical outcomes include quality of life, health status, symptom severity, or other measures that require the patient's participation (e.g., pulmonary function measures, 6-min walk, testing for infection or measures of compliance). While the title of this chapter refers to health-related quality of life (HRQoL), all the issues are relevant to any patient reported outcome. I will for simplicity generally refer only to HRQoL as the outcome measure.

There is not one simple strategy that is appropriate for all studies. A number of reasons complicate the choice of an appropriate design and analysis plan. This chapter provides an overview of the issues that will arise during the design of a trial in the context of two disease settings. While these two examples do not exemplify all possible scenarios, they do provide insight into the thought process that must occur when designing a trial with dropout.

D.L. Fairclough (✉)
Department of Biostatistics and Informatics, Colorado School of Public Health and Colorado Health Outcomes Program (COHO), School of Medicine, University of Colorado Denver
e-mail: Diane.Fairclough@ucdenver.edu

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7_8,

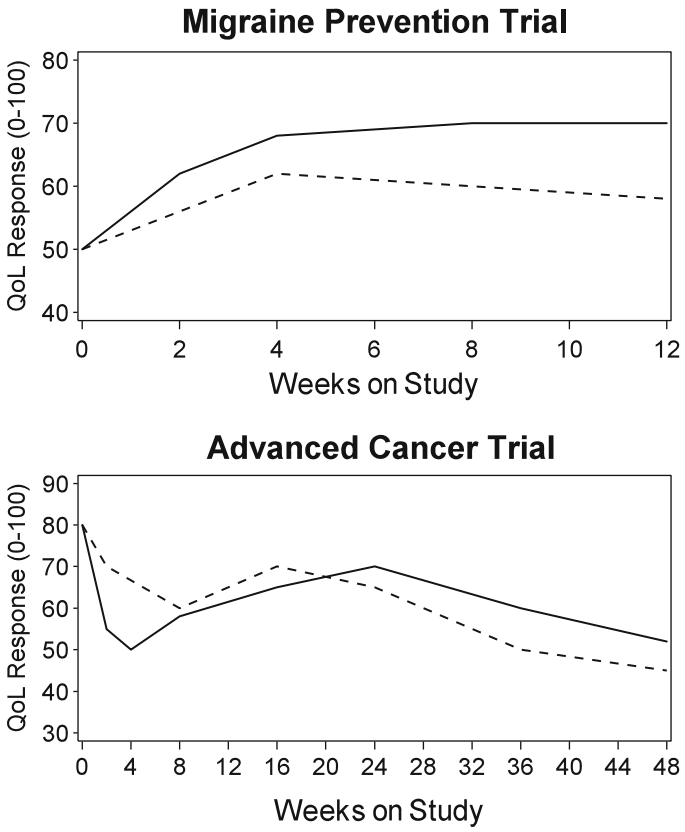

Fig. 8.1 Expected trajectories for two example trials. The *solid line* indicates the experimental arm of each trial

1.1 Examples

1.1.1 Migraine Prevention Trial

The first example is a placebo controlled trial for individuals with frequent migraines intended to reduce the frequency and severity of migraine attacks. During the first month (the titration period), the dose of the experimental drug is slowly increased to minimize side effects. In the second 2 months (the maintenance period), the expected impact of the drug will be measured. In addition to measures of the frequency and severity of migraine episodes, assessments will include a measure of the impact of migraines with three components (HRQoL measures). The expected trajectory of the HRQoL response is illustrated in Fig. 8.1 (left). The response is expected to increase through out the titration period as a result of both a true effect as well as a placebo effect. During the maintenance period, it is expected that the

scores will plateau in the active treatment arm and decrease slightly in the placebo arm. Dropout may be related to the patient's perceived lack of efficacy (which is likely to be related to the outcome) or to side effects (which may or may not be related).

1.1.2 Advanced Cancer Trial

The second example is a trial of patients with advanced cancer comparing standard therapy with standard therapy plus an experimental drug. The experimental drug is expected to delay disease progression but result in more side effects during the early phase of treatment. The primary outcome will be the time to disease progression. Health-related quality of life measures have been added to the trial to assess the impact of both toxicity and disease burden on the patients. The expected trajectories are displayed in Fig. 8.1 (right). Dropout occurs when either the toxicity of treatment is intolerable or there is progressive growth of the tumor indicating that treatment is no longer effective. In both cases, one would expect measures of quality of life, health status, and symptom burden to be worse in subjects who can no longer continue on the trial.

2 Initial Steps

2.1 Aims, Goals, and Audience

Defining the goals and aims of the HRQoL component of a trial is a critical step in the design of the trial, but one of the most challenging steps. It is often the role of the statistician to focus the research team on this task as the information drives the design and analysis components of the trial. This task may be easiest when the HRQoL measures are the primary endpoint. For example, in the advanced cancer trial, the primary aim of the trial may be to extend the time to progression of disease or death. The *aim* of the HRQoL component may be to demonstrate that the more aggressive therapy does not achieve the survival benefits at the cost of the patient's HRQoL or it may be to demonstrate a concurrent HRQoL benefit. In the migraine prevention trial, the primary aim of the trial may be to reduce the frequency and severity of migraines with the goal of the HRQoL component to demonstrate additional benefits as a consequence of the reduced frequency and severity. In both these examples, if the primary aims are not attained then the HRQoL endpoints are not of clinical interest. This suggests that gatekeeping strategies in which families of hypotheses are tested sequentially may be appropriate for controlling the type I error rates (Dmitrienko et al. 2003; Hommel et al. 2007; Fairclough 2009).

There is also a step of explicitly defining the trial endpoints. It is not sufficient to propose a hypothesis such as "The experimental treatment *X* will improve HRQoL compared to standard therapy *C*." Which domains of the HRQoL measure are relevant? Is the focus on the difference at a particular time point or is the interest in the entire trajectory of change? In the advanced cancer trial, one might define the HRQoL endpoint as the average difference between the two treatment trajectories over the first 3 months as a measure of overall HRQoL. In the migraine prevention trial, the focus is on the expected benefit after the drug has been titrated to its full dose. Thus, one might focus on the 12-week assessment or possibly the average of the assessments after 4 weeks ignoring the 2-week assessment.

The goals of a trial go beyond simple statements of aims and general hypotheses. The *goal* may be to inform clinical practice, support drug approval, or obtain marketing claims. In the first goal, the analysis may focus on understanding the trajectories of patients within treatment groups or other subgroups (e.g., survivors). In contrast, the analyses for the second and third goals focus on comparisons between the regimens. Dropout is a threat to both of these goals. The analysis strategies may, however, differ.

2.2 Prevention of Missing Data

Clinical investigators may not fully understand the bias that arises from missing data, so the statistician may need to raise the issue and promote strategies to reduce the prevalence of missing data. There are a number of steps that can be taken during the initial design phase. I divide these steps into primary and secondary prevention. Primary prevention includes steps to prevent or minimize dropout. These steps include having very clear instructions about when and how to approach the participant, what assistance is allowed when there are literacy or language barriers, how to follow-up when an assessment is missed or the participant stops treatment. It is important to be explicit about how long after the targeted time of assessment an assessment is allowed. More relaxed criteria will reduce missing data, but may have less clear interpretations. State the procedure to be followed when treatment is stopped early. Dropout may also be affected by modes of data collection, the design of data collection forms, and training of study personnel.

Secondary prevention involves obtaining auxiliary information that can help with the characterization of the missing data and sensitivity analyses. It is important to know why dropout occurred and if it is related to the outcome(s) of interest. The first task is to make sure that the reason for dropout is collected and the second is to identify wording that will help classify dropout as related or unrelated to the treatment or disease. In the migraine trial, it would be helpful to know if the participant is dropping out due to lack of efficacy, side effects, or for unrelated reasons (moving away, problems with transportation). Auxiliary information takes many forms which may be tailored to the analysis strategy. In a trial where the

participant may become unable to complete the assessment (e.g., Alzheimer's Disease), it may be of value to obtain concurrent information from a caregiver. Adding a very brief assessment of 3–5 key questions between major assessment points may be a possibility; it is also important to have a strategy for attempting to convert refusals to partially completed assessments. In some cases, administrative data (e.g., pharmacy records) may be correlated with the outcomes. Use of auxiliary information is discussed further in later sections.

During the planning of the study, the following questions are typical of those that should be addressed.

- What are the likely causes of attrition or dropout?
- When do you expect attrition to begin and what is the likely rate over the course of the study?
- What patient factors, including the outcome measures, do you expect to be most strongly related to dropout?
- Are there other sources of information that are associated with the HRQoL outcomes that can be collected after the HRQoL assessments are discontinued?

2.3 Timing of Assessments

The timing of assessments is an important design issue in studies with potential dropout. First, we wish to balance participant burden with the need to adequately characterize the trajectory of the outcome over time. Generally, this requires more frequent assessments during the early phase when there is typically more change in the outcome due to the initial response to therapy. More frequent assessments, especially in the early phase, may increase the likelihood of obtaining changes related to the reason for dropout thus making the assumption that follow-up data is missing a random conditional on the already observed data more credible. Consider the migraine prevention trial in which the initial proposal is to assess the patient at baseline and then every 4 weeks for 3 months. If dropout is anticipated in the first 4 weeks, adding an additional assessment at 2 weeks may help to differentiate those who are dropping out due to nonresponse from other reasons.

A final issue is when to discontinue assessments. This is particularly relevant in the advanced cancer trial. If at 6 months, you expect that roughly a third of the patients will have died, it is not useful (and may raise ethical concerns for an Institutional Review Board) to continue to collect HRQoL assessments unless there is a very specific reason clearly identified in the protocol aims. There is no magic rule to determine exactly where to make that cutoff, but some common sense will identify a point beyond which the data will become too sparse to provide reliable estimates. Exploratory analyses of the impact of dropout on power (see Sect. 5) are likely to be informative and help provide cogent arguments to your colleagues.

3 Longitudinal Studies

3.1 Repeated Measures vs. Growth Curve Models

There are two major strategies for analyzing data from longitudinal studies: repeated measures models and mixed effects/growth curve models. This is relevant to studies with dropout because some of the strategies for sensitivity analyses are only applicable to one of these methods. The two approaches can be differentiated based on how time is conceptualized.

In a repeated measures model, time is conceptualized as an ordered categorical variable. This type of design is typical of studies that have distinct phases: pretherapy, during-therapy, and posttherapy. Assessments are classified into distinct ordered categories. If assessments can be classified into distinct windows around the planned assessment times (0, 4, 8, and 12 weeks) one might also adopt a repeated measures model. It important to note that if two or more assessments are classified into any one of the ordered categories all but one will need to be excluded from the analysis or the definition of the categories modified. The models have the general form:

$$Y_{it} = X_{it}\beta + \varepsilon_{it}, \quad \text{Var}[Y_{it}] = \text{Var}[\varepsilon_{it}] = \Sigma_i, \tag{8.1}$$

where i indicates the ith subject and t the tth assessment (ordered category of time). X_{it} indicates the covariates measured on the ith subject at the tth assessment, including indicators of the ordered categories of time. Σ_i has a general (unstructured) form that allows heteroscedasticity over time and different correlations of each pair of assessments on the ith subject. Analysis of repeated measures models using maximum likelihood estimation is widely available in most statistical software packages: SAS and SPSS Mixed procedures and the R gls function in the nlme module.

In mixed effects/growth curve models, time is conceptualized as a continuous variable. Generally, these studies have three or more assessment times, all of which can be included in the analysis regardless of how close they are to other assessments. These models are particularly appropriate when treatment is continuous, without distinct phases, as in the advance cancer trial. The models have the general form:

$$Y_{it} = X_{it}\beta + Z_{it}d_i + \varepsilon_{it}, \tag{8.2}$$

$$\text{Var}[Y_{it}] = \text{Var}[Z_{it}d_i + \varepsilon_{it}] = Z_{it}GZ'_{it} + \mathbf{V}_i, \tag{8.3}$$

where i indicates the ith subject and t the tth assessment. X_{it} indicates the covariates that define the average trajectory or fixed effects and Z_{it} indicates the covariates that define between the subject variation and random effects. The variance of the random effects, G, typically has a general (unstructured) form and the variance of the residual errors typically is assumed to be homoscedastic and uncorrelated,

$\mathbf{V}_i = \sigma_e^2 I$. Analyses of mixed effects models using maximum likelihood (restricted maximum likelihood) estimation is widely available in most statistical software packages: SAS and SPSS Mixed procedures, R lme function in nmle module, and Stata xtmixed.

3.2 Summary Measures

Measures that summarize information across time are particularly useful in longitudinal studies, especially when there are multiple outcome variables. They reduce the multiple comparisons problem and often facilitate interpretation. If the migraine trial has four follow-up assessments, one during the dose titration period (2 weeks) and three during the maintenance period (4, 8, and 12 weeks), then there are potentially four separate tests for each of the outcome measures. However, if we expect the effect of treatment to plateau and specify that our measure of effect is the average of the assessments during maintenance minus baseline, we now have a single measure and test. Alternatively, we might expect it to take the entire 12 weeks to see the benefit that would reflect long-term usage of the medication and define the summary measure to be the change from baseline to 12 weeks. In the advanced cancer trial, we could reduce the number of tests for the 5 follow-up assessments at 2, 8, 16, 24, and 32 weeks to either the average rate of change if we expected the change to be linear or to the area under the curve (AUC) if we expected nonlinear changes.

There are two ways to compute these summary measures. The first is to express them as a linear combination of the parameters for the analysis of the longitudinal data, $\hat{\theta} = C\hat{\beta}$. All of the summary measures specified above would then be estimable as a function of the estimates of β. This has the advantage that if the estimates of β are unbiased, estimates of these summary measures are unbiased and specific rules for handling dropout only need to be identified in the context of the model. The other strategy is to define a function of the individual measures, $\theta_i = f(Y_{it})$ and then perform a univariate analysis. The advantage of the latter approach is that the analysis of the univariate measures is simple and the results may be more interpretable for clinicians. The challenge is in setting up the rules for handling cases with dropout and assessing the impact of various missing data mechanisms. In the migraine study, for the first proposed measure, one might propose that $f(Y_{it})$ is the average of the available maintenance assessments (4, 8, and 12 weeks) minus baseline. This ignores the 2-week data and excludes anyone who drops out early from the analysis. For the second measure (the difference at 12 weeks), everyone not completing the study would be excluded. Similarly, in the advanced cancer trial, estimates of slope would only be available for those with at least two measures and estimates of the AUC would require some method of accounting for the time on study for all patient who dropout. Identifying the rules for all cases a priori and justifying the choices is more challenging than is initially apparent.

4 Missing Data and Options for Sensitivity Analyses

Little and Rubin (1987) provided a classification of missing outcome data based on the dropout process. The dropout process is *missing completely at random* (MCAR) when dropout is independent of both the observed ($Y_i^{obs}, Y_i^{aux}, X_i$) and unobserved ($Y_i^{mis}$) outcome data. It is *missing at random* (MAR) if dropout is independent of the unobserved data, given the observed data. Finally, the dropout process is *missing not at random* (MNAR) if dropout depends on the unobserved data after conditioning on the observed data. The important message is that for maximum likelihood (restricted maximum likelihood) and Bayesian estimation methods, if the dropout is MCAR or MAR and all the observed data is included in the estimation procedure, then the missing data is ignorable. This has two implications. First, one should avoid methods that do not use all the observed data (Y_i^{obs}, X_i). Second, if we can attempt to convert a problem to the MAR assumption by augmenting the observed data with auxiliary information (Y_i^{aux}), methods based on likelihood and Bayesian estimation (multiple imputation, mixture models, and shared parameter models) will be unbiased. Unfortunately, it is not possible to prove that missing data are ignorable (MAR vs. MNAR in the context of likelihood estimation). Thus, sensitivity analyses are strongly recommended when nonignorable dropout is a concern.

Missing covariates result in the exclusion of all observations for a subject for time-independent covariates (e.g., gender, race/ethnicity, insurance status) or the exclusion of a particular observation for time-dependent covariates. A parallel logic exists for this type of missing data, but it is possible to test how the outcome measure depends on the missingness of covariates when the outcome measure is observed. When covariates have missing values, the analyst should carefully consider the proportion of subjects/observations that are excluded from the analysis and the role of the covariates in the analysis. Covariates with missing values should be dropped from the analysis unless the proportion of observations/subjects excluded is very small, or the covariate is essential to the test of a hypothesis.

An additional caution concerns the use of generalized estimating equations (GEE). Without any adjustments, the GEE methods are based on the assumption that dropout is MCAR, a more restrictive assumption than that for likelihood-based methods. The adjustments only relax the assumptions to MAR and while theoretically sound are difficult to implement (Rotnitzky et al. 1998; Tsiatis 2006; Kang and Schafer 2007).

4.1 Last Observation Carried Forward

Use of last observation carried forward (LOCF) is still a commonly practiced method of analysis, though its use is being questioned more frequently. For example, the Food and Drug Administration's draft guidance on patient reported outcomes

raises concerns about this approach (Food and Drug Administration 2006). It is often claimed that it is a *conservative* method. However, most studies that utilize this method fail to perform an exploratory analysis (e.g., plots of trajectories by time of dropout) to verify that assumption. In some cases, it actually may lead to increased type I errors by giving an advantage to treatment regimens where patients with early dropout may appear to have a better response. Except in settings where there is a history of trials where the conservative assumption is true, proposing this as an analysis strategy is risky. An additional concern is when LOCF is used in a setting where a true longitudinal analysis is proposed; it is unclear how this strategies influences the estimation of the covariance structure and test statistics.

4.2 Multiple Imputation

Multiple Imputation (Rubin 1987, 1996) (MI) relies on the assumption that data are MAR conditional on observed information: the observed outcome (Y_i^{obs}), patient characteristics (X_i), and auxiliary information (Y_i^{aux}), sometimes referred to as a surrogate outcome. All but the auxiliary information can be incorporated in the repeated measures or mixed effects models of the available data. So, if there is no auxiliary information that is at least moderately correlated with the outcome, the use of MI techniques will not improve on linear mixed models when nonignorable missing data is a concern (Barnard and Meng 1999; Norton and Lipsitz 2001).

Multiple imputation was originally developed for cross-sectional studies and has been adapted to longitudinal studies. Initially, the technique appears straightforward, but when one gets down to the details in a longitudinal study, it quickly becomes complicated. The first decision is which of the numerous MI models to use. The fact that there are multiple choices is a good indication that none are ideal for all settings. Most require a monotonic missing data pattern. This would occur if there was no missing information in the baseline covariates and the assessments over time (including both Y_i^{obs} and Y_i^{aux}) can be ordered in a way that once one of the outcomes was missing, no additional data was obtained. Only the MCMC-based methods relax this requirement but impose a multivariate normal covariance structure (Schafer 1997). Most methods are based on regression techniques and use information about the relationship between observed variables but ignore information about dropout. The exception is approximate Bayesian Bootstrap (ABB) which uses information about the propensity of each variable to be missing but ignores the correlations between variables. The final challenge is how to handle the multiple measures over time especially in the presence of nonmonotonic patterns. For example, in the migraine study, does one impute the missing values separately for each of the outcomes or does one fill in values for all sequentially across time?

4.3 Pattern Mixture Models

Pattern mixture models (Little 1993) are a popular strategy for attempting to address dropout. The basic idea is that p strata are defined based on the dropout patterns, strata specific parameters are estimated within each pattern ($\hat{\beta}^{(p)}$ and $\widehat{\Sigma}^{(p)}$) using the maximum likelihood methods described previously, and finally the overall parameters, $\hat{\beta}$, are calculated as weighted sums of the strata-specific parameters.

$$Y_i|M^{(p)} \sim N(X_i\beta^{(p)}, \Sigma_i^{(p)}) \quad p = 1, \cdots, P \tag{8.4}$$

$$\hat{\beta} = \sum^{P} \hat{\pi}^{(p)}\hat{\beta}^{(p)}, \tag{8.5}$$

where $\hat{\pi}^{(p)}$ is the proportion of individuals in the pth strata. Because the weights are estimates, the variance is obtained by bootstrapping or an approximation based on the delta method: $\text{Var}(\hat{\pi}^{(p)}\hat{\beta}^{(p)}) \approx \hat{\pi}^{2(p)}\text{Var}(\hat{\beta}^{(p)}) + \hat{\beta}^{2(p)}\text{Var}(\hat{\pi}^{(p)})$.

There are two assumptions associated with these models. The first is that within each strata, the missing data are MAR or ignorable. This first assumption identical to the assumption made in linear mixed models, but relaxes the assumption to a subset of the data. The second is that the restrictions required to estimate all the parameters for repeated measures models or the extrapolations used in the growth curve models are reasonable approximations of the truth. This latter problem often leads investigators to pool strata to facilitate estimation, but this brings into question the validity of the MAR assumption within these larger stratum. Unfortunately, the assumptions cannot be tested. It is strongly recommended that the estimated trajectories for each strata be plotted, thus allowing the reader to assess the plausibility of the restrictions.

Specifying the patterns and the restrictions in the study protocol is challenging when the actual trajectories are unknown. This is a critical problem when the trial is intended for approval of new drugs or new indications of approved drugs or when the trial is intended to be the definitive (and unlikely to repeated) investigation of treatment options. One can attempt to anticipate the likely trajectories, but there can be unexpected complications. Even when one can examine the unblinded trajectories, the solutions are not easy to identify. The migraine prevention trial illustrates a case where a pattern mixture model may be feasible. Let us say that we have chosen to define the outcome as the average of the assessments during the maintenance phase (4, 8, and 12 weeks) minus baseline. If we are to have at least one assessment during the maintenance phase, the first step we would need to make is pool strata for subjects for whom the last assessment is the week 0, 2, or 4. Thus, we are assuming that within this group data are MAR. Then, we might specify that the change from baseline is estimated as the average of the available maintenance assessments. Figure 8.2 illustrates the estimates of the observed trajectories. In this case, the prespecified plan would have been reasonable. But as a counter example, let us consider the advanced cancer trial. Figure 8.3 presents the estimated trajectories for patients with less than 5 weeks, 4–25 weeks and greater than 25 weeks of follow-up as indicated by the time to the last HRQoL

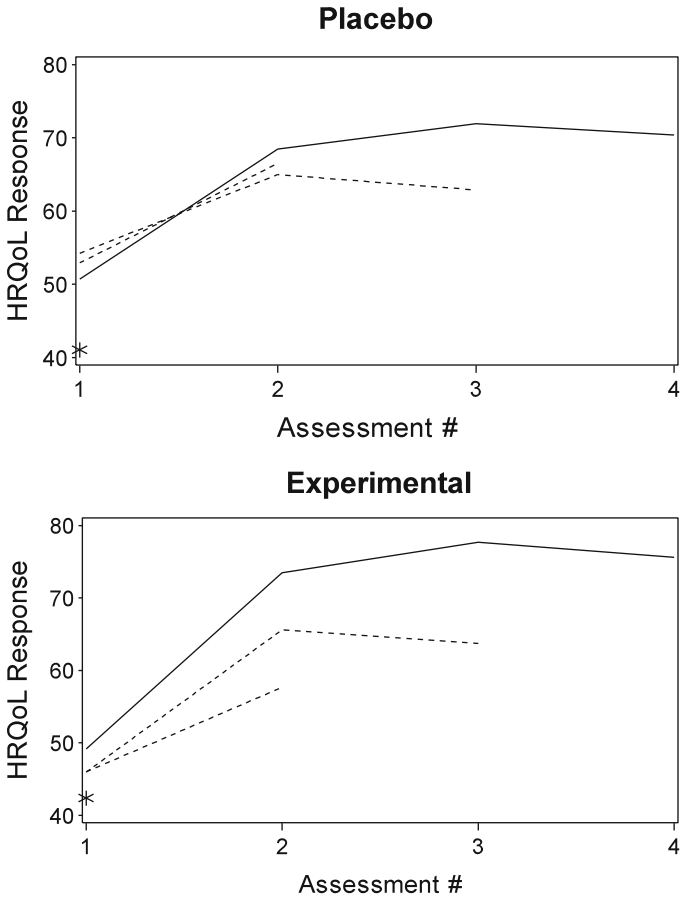

Fig. 8.2 Observed trajectories for the migraine prevention trial in strata defined by length of follow-up for the placebo arm (*left*) and the experimental arm (*right*)

assessment. Even with this information, it would be difficult to identify and defend a strategy to estimate parameters that would characterize the trajectory within each strata and could then be combined across strata. Those with early dropout clearly have a rapid rate of decline, and linear extrapolation over the entire time frame of the trial would likely lead to expected values outside the range of possible scores. A similar problem would arise for the data for patients who dropout between 5 and 25 weeks. A choice to extrapolate the slope just prior to dropout would result in very different projections; the decline over time in the standard therapy arm (left) would seem reasonable, but extrapolating the sharp increase in the experimental arm (right) would be clinically counter intuitive. Daniels and Hogan (2000) suggest an alternative parameterization with centered indicator variables that simplifies the programming.

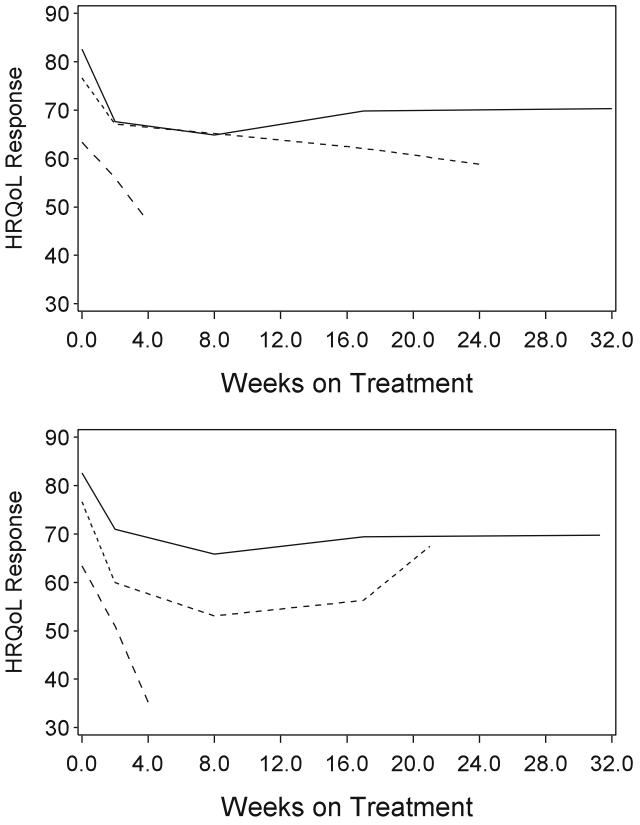

Fig. 8.3 Observed trajectories for the advanced cancer trial in strata defined by length of follow-up for the standard therapy arm (*left*) and the experimental therapy arm (*right*)

4.4 Joint and Shared Parameter Models

An alternative strategy that is applicable in some trials with dropout is to model the outcome of interest and the auxiliary data (Y_i^{aux}) in a joint model. The auxiliary data may be either another longitudinally measured variable, or a variable associated with the dropout mechanism such as time to death. The longitudinally measured auxiliary data could be assessments from a caregiver who can continue to provide evaluations when the patient is no longer able to, a very brief set of key questions that are measured more frequently, or a surrogate measure. In both cases, the hope is that by conditioning on the auxiliary data, the missing data is ignorable (MAR) conditional on the observed and auxiliary data.

The joint and shared parameter models link the outcome of interest and the auxiliary data through the correlation of the data, generally through the random effects. Two equivalent parameterizations have been proposed for models that

include time to an event, such as dropout or death (Schluchter 1992; DeGruttola and Tu 1994; Touloumi et al. 1999). Both use the same general mixed-effects model for the HRQoL outcome:

$$Y_i = X_i\beta + Z_i d_i + e_i. \tag{8.6}$$

They differ in the manner in which *time* is related to the random effects for the HRQoL outcome. In the first alternative, the model for time to the event is

$$f(T_i) = \mu_T + r_i. \tag{8.7}$$

The two models are joined by allowing the random effects (d_i) to covary with the residual errors of the time model (r_i) and thus with time to the event itself.

$$\begin{bmatrix} d_i \\ r_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D & \sigma_{bt} \\ (\sigma_{bt})' & \tau^2 \end{bmatrix} \right). \tag{8.8}$$

$$\begin{bmatrix} Y_i \\ f(T_i) \end{bmatrix} \sim N\left(\begin{bmatrix} X_i\beta \\ \mu_T \end{bmatrix}, \begin{bmatrix} Z_i D Z_i' + \sigma^2 I & Z_i\sigma_{bt} \\ \sigma_{bt} Z_i' & \tau_a^2 \end{bmatrix} \right). \tag{8.9}$$

In the second parameterization, the random effects (d_i) are included in the time to event model:

$$f(T_i) = \mu_T + \lambda d_i + t_i. \tag{8.10}$$

$$\begin{bmatrix} Y_i \\ f(T_i) \end{bmatrix} \sim N\left(\begin{bmatrix} X_i\beta \\ \mu_T \end{bmatrix}, \begin{bmatrix} Z_i D Z_i' + \sigma^2 I & Z_i D\lambda \\ \lambda D Z_i' & \lambda D\lambda' + \tau_b^2 \end{bmatrix} \right). \tag{8.11}$$

The two models are equivalent as the parameters of one can be written as a function of the parameters of the other model: $\sigma_{bt} = \lambda D$ and $\tau_a^2 = \lambda D\lambda' + \tau_b^2$. Both alternatives have specific uses. The first alternative is more intuitive when the focus is on the HRQoL outcome. In contrast, the second alternative expresses the time to the event as a function of the longitudinal random effects. The second alternative allows one to use the `SAS NLMixed` procedures to obtain maximum likelihood (ML) estimates of the parameters. These models have been a popular area of research and this simple model has been extended to alternative models for the time to dropout (Vonesh et al. 2006; Li et al. 2007; Rizopoulos et al. 2009) and competing causes of dropout (Law et al. 2002; Elashoff et al. 2007; Jaros 2008).

The approach is very similar when the outcome of interest (Y_i) and the auxiliary information (Y_i^*) are both measured longitudinally. The trajectories of the observed and auxiliary information are related through the random effects and possibly the residual errors:

$$\begin{aligned} Y_i &= X_i\beta + Z_i d_i + e_i \\ Y_i^* &= X_i^*\beta^* + Z_i^* d_i^* + e_i^* \end{aligned}$$

$$\begin{bmatrix} d_i \\ d_i^* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D_1 & D_{12} \\ (D_{12})' & D_2 \end{bmatrix} \right) \text{ and } \begin{bmatrix} e_{it} \\ e_{it}^* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1 & \sigma_{12} \\ (\sigma_{12})' & \sigma_2 \end{bmatrix} \right). \tag{8.12}$$

The timing of the assessments does not have to be the same for the two measures. The auxiliary information could be measured more frequently than the outcome of interest. As an example, consider a trial where a pulmonary function test or 6-min walk is the primary measure of interest. Some patients, particularly those who are doing poorly, may be unable to perform the test. However, they may be able to answer a short series of questions assessing their ability to perform physical daily activities. If changes in the brief questions, Y_i^*, are correlated with changes in the outcome of interest, Y_i, there will be partial information about the response that can reduce the bias do to the missing assessments.

4.5 Analysis of Blinded Data

The proportion of subjects who drop out and other characteristics of the data drive the need for a sensitivity analysis and the choice of the alternative models. If the proportion of subjects who dropout is small (generally $<5\%$), the missing data will have little impact on the analysis. Other characteristics may drive the choice among the alternative models. For example, if there is minimal variation in the random effects, particularly random effects associated with change over time, the joint and shared parameter models will be difficult to implement as it will be difficult to estimate correlations between entities with little variation. Plots such as Figs. 8.2 and 8.3 would facilitate the implementation of a pattern mixture model. Obviously, this information is not available at the time the analysis plan in the protocol is written. But a strategy that might be considered is to propose the analyses that are anticipated and also propose to examine data pooled across treatment groups before it is unblinded for the rates of dropout, trajectories within groups defined by dropout and variance of the random effects pooled over the treatment groups to see if the proposed sensitivity analyses seem reasonable. While there is still no guarantee that problems will not arise, the investigators can report that any modifications to the analysis plan were made prior to the unblinding of the data, thus less suspect to manipulation.

4.6 HRQoL After Death or Discontinuation of Therapy

The issue of interpreting HRQoL after death is controversial. With the exception of utility (preference) measures,[1] attempting to estimate HRQoL after death does not have a conceptual basis. However, in any analysis that compares HRQoL outcomes across intervention groups, there is either implicit or explicit imputation of these missing values. This is most obvious in the context of the EM algorithm for

[1]These measures are defined as having a value of 1 for perfect health and 0 for death.

maximum likelihood estimation. In the E-step, the sufficient statistics are functions of the observed data and the conditional expectation (imputed values) of the missing data. The point is that missing values after death cannot be entirely ignored in the context of the analysis. Some might argue that any analysis that includes measurements after death is not interpretable. But this same logic would exclude the analysis of any outcome except survival, including measures of symptoms (pain) or physical function (6-min walk), in populations with any mortality. I would suggest that analyses in populations with mortality are useful, but that the analyst needs to pay careful attention to the underlying assumptions and should supplement the primary analysis with additional models that explore the sensitivity to plausible alternatives.

Some of the same concerns exist for trials where the measurement of the outcome is discontinued after therapy is terminated. In the advanced cancer trial, treatment using the trial drugs will be discontinued if progression of the disease (growth of the tumor) or unacceptable toxicity is observed. If the assessment of the outcome is discontinued at that point, similar issues exist with respect to the estimation and interpretation of the future trajectory of those individuals. If the assessment continues, but other treatment is initiated, the interpretation of the results becomes more complex. In the context of the advanced cancer trial, I would argue from an intent-to-treat perspective that assessment should be continued and included in the analysis recognizing that HRQoL is likely to decline due to the symptoms associated with the progressive disease or initiation of even more aggressive and toxic therapy. The same arguments would not apply to the migraine prevention trial, where discontinuation of therapy is likely to be a consequence of toxicity or lack of efficacy which would have negative impact on HRQoL and switching to a new therapy a potentially positive impact. The strategies to address these concerns will vary across settings and will include careful documentation of the causes of dropout, description of trajectories within selected subgroups and sensitivity analysis using plausible assumptions.

5 Sample Size Estimation

Sample size (or power) estimation for longitudinal studies is a challenge. The presence of missing data further complicates the process. None of the routines in the statistical packages that compute sample size address this problem. This is partially due to the number of unknown quantities that need to be specified because of the longitudinal nature of the study. These include the differences between the intervention arms at each of the repeated assessments under the alternative hypothesis as well as all the covariance parameters. When specifying the differences under the alternative hypothesis, one might consider both the expected differences (generally based on pilot data) or various scenarios that would be compelling enough to change clinical practice. The specification of the expected differences will be affected by a number of factors including the expected mechanism of action.

For example, an intervention of limited duration might produce the greatest effect shortly after the intervention with diminishing effect over time. In contrast, another intervention might produce a small effect initially with an increasing difference over time. Complete specification of the covariance structure will require pilot data. When pilot data does not exist, there are possible simplifications including defining the effects and covariance assuming a standard deviation of 1 and a constant correlation of the repeated assessments. For HRQoL and other patient reported measures, the correlations are generally in the range of 0.4–0.8.

Given the lack of routines for sample size calculations in longitudinal studies in statistical packages, what are the options? The first is to choose a simple outcome measure (e.g., the last assessment), calculate the sample size needed pretending that the data would be analyzed using a two sample t-test and then inflate that number to account for dropout by dividing by the proportion of subjects expected to have a final assessment. The procedure is simple, but generally overestimates that required sample size.

The second strategy relies on the specification of an endpoint that is a linear combination of the parameters describing the longitudinal trajectory, such as those described in Sect. 3.2. The procedure involves specifying (1) the hypothesis as a linear function of the parameters ($H_0 : \theta = C\beta - G = 0$), (2) the expected difference under the alternative hypothesis of the outcome measure ($H_a : \delta_\theta = C\beta - G$), (3) the expected rate of dropout Between each assessment, and (4) the correlation or covariance of the assessments over time (Σ). If the sample size is sufficiently large, then the distribution of the test statistic for the hypothesis can be approximated by a univariate normal distribution. For incomplete data, the relationship for a two-sided test with type I and II errors of α and β can be written as

$$(z_{\alpha/2} + z_\beta)^2 = \frac{\delta_\theta^2}{\mathrm{Var}(C\hat{\beta})}, \ \mathrm{Var}(C\hat{\beta}) \approx C \sum_{n=1}^{N} (X_i' \Sigma_i^{-1} X_i)^{-1} C'. \tag{8.13}$$

When missing data occurs solely as a function of dropout, $X_i = X_k$ for Np_k subjects in the same treatment group who dropout between two assessments, where N is the total sample size and $p_k, k = 1, \ldots, K$ is the proportion of subjects with the design matrix X_k. X_k may include indicator variables for the planned assessments or functions of time depending on the planned analysis as described previously in Sect. 3.1. Then, $\sum_{n=1}^{N} X_i' \Sigma_i^{-1} X_i$ can be rewritten as $N \sum_{k=1}^{K} p_k X_k' \widehat{\Sigma}_k^{-1} X_k$. Solving for N yields:

$$N = (z_{\alpha/2} + z_\beta)^2 C \left[\sum_{k=1}^{K} p_k X_k' \widehat{\Sigma}_k^{-1} X_k \right]^{-1} C' / \delta_\theta^2. \tag{8.14}$$

The calculation of $\left[\sum_{k=1}^{K} p_k X_k' \widehat{\Sigma}_k^{-1} X_k\right]^{-1}$ is possible with programs that handle matrix mathematics. A second alternative is possible if you can input the covariance structure into the procedure or function that maximizes the likelihood and stop the

algorithm after the very first iteration as with the SAS Mixed procedure.[2] Other adaptations for tests with multiple degrees of freedom, small sample sizes, and intermittent missing data are possible (Fairclough 2009).

The final option involves simulation of X_i and may be useful when the missing data pattern includes intermittent missing data or the timing of assessments varies across individuals.

5.1 Migraine Example

Consider the Migraine Prevention Trial with five assessments over time. The first step is to identify the endpoint and it's expected value under the alternative hypothesis. The expected differences between the two groups at 0, 2, 4, 8 and 12 weeks are 0, 6, 6, 10, and 12 points, respectively (Fig. 8.1 (left)). Let us assume dropout is strictly monotonic and that 5% of the sample drops out after each assessment, with 80% completing all assessments. Then, we assume that the standard deviation is 20 and the correlation between assessments is 0.5 (a lower bound for most HRQoL measures).

Using the first strategy, if our endpoint is the change from baseline to the 12-week assessment, then we may simplify the problem to the comparison of the 12-week assessments, where δ_θ is 12 points. Using standard software for a two sample test ($\alpha = 0.05$, two-sided test, 90% power), we would need a total of 120 subjects with the 12 week assessment. Inflation for 80% dropout by 12 weeks would require a total of 150 subjects to be enrolled.

If we wish to use the second strategy and a repeated measures analysis (see Sect. 3.1) was planned with a cell mean model, the design matrix, X_i, will consist of indicator variables for each time (t) and treatment arm (k) combination. β will be a vector of the TK corresponding means, μ_{tj}. C is then defined to generate θ. In this simple example, if

$$\beta = (\mu_{0A}\ \mu_{2A}\ \mu_{4A}\ \mu_{8A}\ \mu_{12A}\ \mu_{0B}\ \mu_{2B}\ \mu_{4B}\ \mu_{8B}\ \mu_{12B})'$$

and $C = (1\ 0\ 0\ 0\ -1\ \ -1\ 0\ 0\ 0\ 1)$, then $\theta = C\beta = (\mu_{12B} - \mu_{0B}) - (\mu_{12A} - \mu_{0A})$ or the treatment difference in the change from baseline to 12 weeks. With monotone dropout, there will be ten unique design matrices (five for each treatment arm). Table 8.1 summarizes the expected proportions with each pattern of dropout and the corresponding value of p_k for a 1:1 and 1:2 allocation to treatment group A and B. Utilizing all the data, assuming equal allocation and using the asymptotic approximation, the total sample size is 136 subjects, roughly a 10% reduction compared to the first strategy. If the correlation of assessments over time is increased ($\rho = 0.6$), the sample size is reduced to 110 subjects. For an alternative endpoint (average of

[2]Examples are available on http://home.earthlink.net/~dianefairclough/Welcome.html.

Table 8.1 Proportion of subjects (p_k) with unique design matrices for 1:1 and a 1:2 allocation to treatment group A and B

		Week of assessment					Allocation	
Pattern	Group	0	2	4	8	12	1:1	1:2
1	A	X					0.025	0.0167
2	A	X	X				0.025	0.0167
3	A	X	X	X			0.025	0.0167
4	A	X	X	X	X		0.025	0.0167
5	A	X	x	X	X	X	0.400	0.2667
6	B	X					0.025	0.0333
7	B	X	X				0.025	0.0333
8	B	X	X	X			0.025	0.0333
9	B	X	X	X	X		0.025	0.0333
10	B	X	x	X	X	X	0.400	0.5333

4, 8, and 12 weeks – baseline), $C = (10 - 1/3 - 1/3 - 1/3 - 101/31/31/3)$ and $\delta_\theta = (6 + 10 + 12)/3 - 0 = 9.33$. The sample size estimates are 140 and 112 for $\rho = 0.5$ and 0.6, respectively. Given that we are at best making educated guesses about the dropout rate, the expected trajectories and the correlation, it would be wise to examine a number of likely scenarios.

5.2 Advanced Cancer Trial

As a second example, let us assume that the sample size for the trial has been determined by the primary endpoint to be a total of 300 patients. The question of interest is what is the power to detect a clinically relevant difference (say $\delta_\theta = 5$ points) in the change from baseline to the end of the study for the primary HRQoL measure. We are specifically interested in the last three assessments, 24, 36, and 48 weeks because we expect roughly half of the subjects would have progressed by 48 weeks. Also assume that the standard deviation is expected to be 15 points and the correlation of assessments over time to be approximately 0.6. The proportion with HRQoL assessments is projected to be 100, 95, 09, 80, 70, 50, 30% at 0, 2, 8, 12, 24, 36, and 48 weeks. The power to detect the difference of interest is 0.81, 0.72, and 0.55 at 24, 36, and 48 weeks, respectively. If the correlation among assessments over time is 0.5, the power is reduced to 0.73, 0.62, and 0.47 respectively. The results suggest a diminishing value of assessments after 24 weeks if the criteria is solely based on the power to detect the hypothesized difference.

6 Summary

Designing a clinical trial in which more than a small rate of dropout is expected is very challenging. The issues are the same for most patient-reported outcomes as well as other measures that require the study subject direct participation (pulmonary function tests, 6-min walk, grip strength). Early involvement of the statistician increases the likelihood that the analytic challenges and barriers in interpretation of results can be reduced.

Clearly, prevention of missing data is the primary goal, but in many settings (e.g., dropout due to death), it is not possible to eliminate the problem. When every effort to eliminate preventable missing data due to administrative causes has been implemented, it will be rare that one does not have some suspicion that dropout in most individuals is related to the outcomes of interest and thus clearly not MCAR and possibly not MAR. Analysis plans should employ methods that make the least restrictive assumptions (thus avoiding all methods that assume MCAR). In most cases, the best strategy will be to employ likelihood based/Bayesian methods that utilize all the available data as the primary analysis, supplemented by sensitivity analyses using pattern mixture models, joint/shared parameter models or multiple imputation to assess the sensitivity of the results to alternative assumptions. There is a common theme underlying all of these alternative models; specifically, each in some manner tries to convert the analysis to one where conditional on either auxiliary data or the specification of strata, the data can be considered MAR. The choice between the various methods will depend on the availability and form of the auxiliary data; so it is critical that strategies for obtaining auxiliary data are incorporated into the design of the study.

References

Barnard J, Meng XL (1999) Applications of multiple imputation in medical studies: From AIDS to NHANES. Stat Meth Med Res 8:17–36

Daniels MJ, Hogan JW (2000) Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. Biometrics 56:1241–1248

DeGruttola V, Tu XM (1994) Modeling progression of CD-4 lymphocyte count and its relationship to survival time. Biometrics 50:1003–1014

Dmitrienko A, Offen W, Westfall PH (2003) Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Stat Med 22:2387–2400

Elashoff RM, Li G, Li N (2007) An approach to joint analysis of longitudinal measurements and competing risks failure time data. Stat Med 26:2813–2835

Fairclough DL (2009) Design and analysis of quality of life studies in clinical trials, 2nd edn. Chapman & Hall, FL

Food and Drug Administration (2006) Draft Guidance for Industry on Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims, 71 Federal Register 5862. February 3, 2006

Hommel G, Bretz F, Maurer W (2007) Powerful sort-cuts for multiple testing procedures with special reference to gatekeeping strategies. Stat Med 26:4063–4073

Jaros M (2008) A joint model for longitudinal data and competing risks. Doctoral dissertation, University of Colorado at Denver

Kang DY, Schafer JL (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion and rejoinder). Stat Sci 22:523–539

Law NJ, Taylor JMG, Sandler HM (2002) The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. Biostatistics 3:547–563

Li L, Hu B, Greene T (2007) Semiparametric joint modeling of longitudinal and survival data. Joint Statistical Meetings Proceedings 2007, pp 293–300

Little RJ, Rubin DB (1987) Statistical analysis with missing data. Wiley, NY

Little RJA (1993) Pattern-mixture models for multivariate incomplete data. J Am Stat Assoc 88:125–134

Norton NJ, Lipsitz SR (2001) Multiple imputation in practice: Comparison of software packages for regression models with missing variables. Am Stat 55:244–254

Rizopoulos D, Verbeke G, Lesaffre E (2009) Fully exponential Laplace approximation for the joint modelling of survival and longitudinal data. J Roy Stat Soc Series B 71:637–654

Rotnitzky A, Robins JM, Scharfstein DO (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. J Am Stat Assoc 93:1321–1339

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

Rubin DB (1996) Multiple imputation after 18+ years. J Am Stat Assoc 91:473–489

Schafer J (1997) Analysis of incomplete multivariate data. Monograph on Statistics and Applied Probability 72. Chapman and Hall, London

Schluchter MD (1992) Methods for the analysis of informatively censored longitudinal data. Stat Med 11:1861–1870

Touloumi G, Pocock SJ, Babiker AG, Daryshire JH (1999) Estimation and comparison of rates of change in longitudinal studies with informative drop-outs. Stat Med 18:1215–1233

Tsiatis AA (2006) Semiparmetric theory and missing data. Springer, New York

Vonesh EF, Greene T, Schluchter MD (2006) Shared parameter models for the joint analysis of longitudinal data and event times. Stat Med 25:143–163

Index

D. Harrington (ed.), *Designs for Clinical Trials: Perspectives on Current Issues*, Applied Bioinformatics and Biostatistics in Cancer Research, DOI 10.1007/978-1-4614-0140-7,