Springer Series in Statistics

Sequential Experimentation in Clinical Trials

Design and Analysis

Springer

Springer Series in Statistics

For further volumes:
http://www.springer.com/series/692

Jay Bartroff • Tze Leung Lai • Mei-Chiung Shih

Sequential Experimentation in Clinical Trials

Design and Analysis

Springer

Jay Bartroff
Department of Mathematics
University of Southern California
Los Angeles, California, USA

Tze Leung Lai
Department of Statistics
and Cancer Institute
Stanford University
Stanford, California, USA

Mei-Chiung Shih
Department of Health Research and Policy
Stanford University
VA Cooperative Studies Program
Stanford, California, USA

ISSN 0172-7397
ISBN 978-1-4614-6113-5 ISBN 978-1-4614-6114-2 (eBook)
DOI 10.1007/978-1-4614-6114-2
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012951627

Printed on acid-free paper

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

To our parents

Jack and Barbara Bartroff

Chi Y. and Wai C. Lai

Ming-Tsan Shih and Min-Hui Hsu

Preface

The idea of writing a book on sequential experimentation in clinical trials arose 25 years ago when Lai was at Columbia University, collaborating with Dan Anbar of Abbott Laboratories on a university-industry cooperative research project, “Sequential Statistical Methods in Biopharmaceutical Research,” funded by the National Science Foundation. Anbar and Lai, together with Gordon K.K. Lan and Anastasio Tsiatis (at that time, at the NIH and Harvard, respectively), formed a focused research group that held a week-long meeting every 2 months and organized an annual workshop, with invited speakers from academia, industry, and the FDA and NIH, and dedicated to the development and discussion of sequential methods in the design and analysis of clinical trials. Although substantial progress was made by the group to advance this new area that attracted considerable attention from the pharmaceutical industry after the early termination of the Beta-Blocker Heart Attack Trial in 1981, the book project could not materialize when Lai moved to the West Coast in 1987, joining Stanford University, while the other collaborators remained on the East Coast but were busy with their own moves to new positions. On the other hand, the group members continued their separate research efforts in this area. These efforts and those by other researchers led to major advances and eventual widespread use of group sequential designs and interim analysis methods by the pharmaceutical industry and their acceptance by the FDA.

At the turn of the new century, the monograph by Jennison and Turnbull (2000) appeared, giving a comprehensive overview of group sequential methods developed up to that time. Besides continual developments in interim analysis and group sequential methods, the past decade has also witnessed new developments and growing interest in adaptive designs of clinical trials. The books by Proschan et al. (2006), Chow and Chang (2006), Chang (2007), and Berry et al. (2011) describe some of these developments and their applications. However, as pointed out in Chap. 8, there is substantial disagreement in the literature concerning the appropriateness of these adaptive designs, which either use inefficient test statistics that are not supported by mainstream statistical principles to adjust for the adaptation in maintaining the Type I error of the test or use Bayesian posterior probabilities that do not guarantee the prescribed Type I error. Chapter 8 describes our recent work that provides a new class of adaptive designs which are both flexible and efficient, thereby resolving the dilemma between efficiency and flexibility in the adaptive design literature. Prior to this work, we have also developed a comprehensive methodology of flexible and efficient group sequential designs, to which Chap. 4 is devoted. In fact, the new adaptive designs in Chap. 8 are modifications of the corresponding group sequential designs in Chap. 4, and a unified approach is provided for the methodology and implementation of group sequential and adaptive designs.

Besides giving an up-to-date account of these flexible designs, we also present in Chap. 7 a comprehensive overview, including the most recent developments of inference after the termination of these clinical trials. Chapter 6 describes the Beta-Blocker Heart Attack Trial as an example for the design and analysis of clinical trials with failure-time endpoints and interim analyses. The material in Chaps. 4, 6, 7, and 8 can be used for short courses on group sequential and adaptive designs. We have given short courses based on this material in the First Joint Biostatistics Symposium in Beijing, July 2010, the Applied Statistics Symposium of the International Chinese Statistical Association in New York, June 2011, and the Workshop on the Design and Analysis of Clinical Trials at National University of Singapore, October 2011. We were greatly encouraged by the enthusiastic response and stimulating comments of the participants.

This book has also benefited from the Third International Workshop in Sequential Methodologies at Stanford University, June 2011. The workshop was very well attended and was truly international in nature. There it was pointed out that despite a resurgence of interest in sequential analysis, the subject was not in the graduate curriculum of most statistics departments. One reason that was mentioned was the lack of textbooks that could present the material in an appealing way to today's graduate students. In fact, only a handful of such books had been written and they were published more than 20 years ago. Although there are more recent books which we have mentioned in the second paragraph of this preface, they all deal with the specialized topics of group sequential and adaptive designs rather than general methods and principles in sequential analysis. Another reason that came up during workshop discussions was that sequential methods and adaptive designs seemed to involve special techniques and ideas that are detached from mainstream topics taught in the modern graduate statistics curriculum, e.g., likelihood inference, regression analysis, resampling, semiparametric theory, to name a few. Spurred by these comments, we have made particular efforts to change this perception in the selection and presentation of the materials. To make it suitable for an introductory course on sequential analysis, the book covers the much broader subject of sequential experimentation that includes group sequential and adaptive designs of Phase II and III clinical trials, which have attracted much attention in the past three decades. In particular, the broad scope of design and analysis problems in sequential experimentation clearly requires a wide range of statistical methods and models from nonlinear regression analysis, experimental design, dynamic programming, survival analysis, resampling, and likelihood and Bayesian inference. The background material in these building blocks is summarized in Chaps. 2 and 3 and certain sections in Chaps. 6 and 7. Besides group sequential tests and adaptive designs, we also introduce sequential change-point detection methods in Chap. 5 in connection with pharmacovigilance and public health surveillance. Together with dynamic programming and approximate dynamic programming in Chap. 3, the book therefore covers all basic topics for a graduate course in sequential analysis.

Different parts of the book can be used for short courses on clinical trials, translational medical research, and sequential experimentation. Lai has used an early draft of the book to teach a course on innovative clinical trial designs and statistical methods for second-year Ph.D. students in the Department of Statistics at Stanford University. The course has led to supplements and exercises for various chapters and also to the web site for the book, http://meichiun.web.stanford.edu/clinicaltrials/, to which different parts of the book refer for links to software.

We thank the teaching assistant Alex Deng and the students in the aforementioned course for their stimulating interest and helpful comments on a first draft of the book that was used as lecture notes. In particular, we want to mention Olivia Liao who subsequently wrote her Ph.D. thesis on the subject and contributed to some recent publications that we have referenced. We also thank our collaborators Philip Lavori, Zheng Su, Ray Zhu, Joe Heyse, Jie Chen, Sam Wong, and Hock Peng Chan for working with us on related projects and for their valuable suggestions. We are particularly grateful to Balasubramanian Narasimhan who has been helping us develop open-source software for the innovative designs and analyses that will be posted at the book's web site and to Cindy Kirby who has put together the separate chapters in an efficient and timely fashion. We also acknowledge grant support for the research projects related to this book by the National Science Foundation (DMS-0907241 and 0403105 for Bartroff and 1106535 and 0805879 for Lai), the National Institutes of Health (GMS-068968 for Bartroff and 1 P30 CA124435 for Lai), the U.S. Department of Veterans Affairs Cooperative Studies Program (for Shih), and the Stanford Center for Clinical and Translational Education and Research (for Lai and Shih).

Los Angeles, California Jay Bartroff
Stanford, California Tze Leung Lai
Stanford, California Mei-Chiung Shih

Contents

Chapter 1
Introduction

This chapter gives an overview of (a) the prevalence of sequential experimentation in translational medical research and (b) developments of statistical methods to design and analyze these sequential experiments in evidence-based medical research. In this connection it also gives an outline of the topics covered in the subsequent chapters and discusses the complementary roles of Bayesian and frequentist approaches to sequential design and analysis.

1.1 Sequential Experimentation in Translational Medical Research

"From bench to bedside," a maxim of translational medical research, reflects the sequential nature of the experiments involved. "Bench" refers to laboratory experiments to study new biochemical principles and discover novel treatments. The experiments with promising results are followed by preclinical animal studies. After understanding the effect of the treatment (say, a new drug) on animals (e.g., rodents), the next stage of drug development consists of clinical trials that involve human subjects, starting with Phase I studies to determine a safe dose or dosage regimen and to collect information on the pharmacokinetics (PK) and pharmacodynamics (PD) of the drug. PK is concerned with the concentration versus time curve that is associated with the kinetics of drug absorption, distribution, and elimination. PD is concerned with the steady-state relationship of drug concentration at an effector site to the effect/response produced. The information collected and the dosage regimen determined from Phase I studies are used to design Phase II clinical trials to evaluate the efficacy of the drug for particular indications (endpoints) in patients with the disease. Phase II trials are precursors of Phase III trials whose goal is to demonstrate effectiveness of the drug for its approval by the regulatory agency (the Food and Drug Administration in the United States) and to provide adequate evidence for its labeling. Besides testing efficacy, Phase III trials also collect safety information

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_1,

from the relatively large samples of patients accrued to the trial. The safety of the drug is evaluated from the data obtained from all three phases of clinical trials prior to marketing approval of the drug and continues to be evaluated through post-marketing Phase IV trials.

Despite the sequential nature of Phase I–III trials, the trials are often planned separately, treating each trial as an independent study whose design depends on results from studies in previous phases. An advantage of this is that the reproducibility of the results of the trial can be evaluated on the basis of the prescribed design, without worrying about the statistical variability of the results of earlier-phase trials that determine the prescribed design. A disadvantage lies in the fact that the sample sizes of the trials are often inadequate because of the separate planning. A different strategy is to expand a trial seamlessly from one phase into the next phase; the Phase II–III cancer trial design in Sect. 6.7 is an example. Although Phase II–III design, which is an active area of current research undergoing new advances, is beyond the scope of this book, we give a brief introduction in Sect. 6.7 to show the power of an overarching sequential experimentation approach to translational medicine.

This book focuses on sequential methods for the design and analysis of Phase I, II, and III clinical trials, thereby providing the background for understanding and developing the new advances. Although these methods are developed in the context of clinical trials, they are also applicable to other fields that involve sequential experimentation. We therefore give an introduction to the statistical methods and the underlying principles and also relate them to basic topics taught in typical graduate statistics programs that assume the data to be generated by nonsequential designs. For example, while Chap. 2 considers Phase I clinical trials, it starts with nonlinear regression and experimental design before relating them to basic pharmacologic principles and models underlying dose determination.

1.2 Sequential Analysis: From Weapons Testing to Confirmatory Clinical Trials

The subject named *sequential analysis*, which also includes sequential design of experiments, was born in response to demands for more efficient testing of antiaircraft gunnery during World War II, which led to Wald's development of the sequential probability ratio test (SPRT) in 1943 (Wallis 1980). Let $X_1, X_2, \ldots$ be i.i.d. random variables with common density function f. To test $H_0 : f = f_0$ versus $H_1 : f = f_1$, the SPRT stops sampling at stage

$$N = \inf\left\{ n \geq 1 : \prod_{i=1}^{n} \big(f_1(X_i)/f_0(X_i) \big) \notin (A, B) \right\}, \tag{1.1}$$

where $0 < A < 1 < B$ are the stopping boundaries. When stopping occurs, H_0 or H_1 is rejected according to whether the likelihood ratio $\prod_{i=1}^{N} (f_1(X_i)/f_0(X_i))$ crosses the upper boundary B or the lower boundary A. In Chap. 3 we give a summary of the theory of sequential tests of hypotheses, beginning with the SPRT on testing a simple null versus a simple alternative hypothesis and describing important subsequent developments that led to a relatively complete theory for composite hypotheses.

Within a few years after Wald's introduction of the SPRT, it was recognized that sequential hypothesis testing might provide a useful tool in clinical trials to test the efficacy of new medical treatments. A number of papers appeared during the 1950s on modifications of the SPRT for the design of clinical trials, and an overview of these developments was given in Armitage (1960). In 1969, Armitage et al. proposed a new alternative to the SPRT and its variants, called the *repeated significance test* (RST). The underlying motivation for the RST is that, since the strength of evidence in favor of a treatment from a clinical trial is conveniently indicated by the results of a conventional significance test, it is appealing to apply the significance test, with nominal significance level α, repeatedly during the trial. Noting that the overall significance level α^*, which is the probability that the nominal significance level is attained at some stage, is larger than α, they developed a recursive numerical algorithm to compute α^* in the case of testing a normal mean θ with known variance σ^2, for which the RST of $H_0 : \theta = 0$ is of the form

$$T = \inf\left\{n \le M : |S_n| \ge a\sigma\sqrt{n}\right\}, \tag{1.2}$$

rejecting H_0 if $T < M$ or if $T = M$ and $|S_M| \ge a\sigma\sqrt{M}$, where $S_n = X_1 + \cdots + X_n$. Haybittle (1971) proposed the following modification of the RST to increase its power. The stopping rule has the same form as (1.2) but the rejection region is modified to $T < M$ or $|S_M| \ge c\sigma\sqrt{M}$, where $a(\ge c)$ is so chosen that the overall significance level is equal to some prescribed number. In particular, $a = \infty$ gives the fixed sample size test while $a = c$ gives the RST.

In double-blind multicenter clinical trials, it is not feasible to arrange for continuous examination of the data as they accumulate to perform the RST. This led Pocock (1977) to introduce a "group sequential" version of (1.2), in which the X_n represents an approximately normally distributed statistic of the data in the nth group (instead of the nth observation) and M represents the maximum number of groups. Instead of the square-root boundary $a\sigma\sqrt{n}$, O'Brien and Fleming (1979) proposed to use a constant stopping boundary in

$$T = \inf\{n \le M : |S_n| \ge b\}, \tag{1.3}$$

which corresponds to the group sequential version of an SPRT.

While sequential analysis had an immediate impact on weapons testing when it was introduced during World War II to reduce the sample sizes of such tests (Wallis 1980), its refinements for testing new drugs and treatments received little attention from the biomedical community until the Beta-Blocker Heart Attack Trial (BHAT) that was terminated in October 1981, prior to its prescheduled end in June 1982.

The main reason for this lack of interest is that the fixed sample size (i.e., the number of patients accrued) for a typical trial is too small to allow further reduction while still maintaining reasonable power at the alternatives of interest. On the other hand, BHAT, which was a multicenter, double-blind, randomized placebo-controlled trial to test the efficacy of long-term therapy with propranolol given to survivors of an acute myocardial infarction, drew immediate attention to the benefits of sequential methods not because it reduced the number of patients but because it shortened a 4-year study by 8 months, with positive results for a long-awaited treatment for MI patients.

The "success story" of BHAT paved the way for major advances in the development of group sequential methods in clinical trials and for the steadily increasing adoption of group sequential design. Chapter 4 gives a review of these advances and describes the current methodology that has moved far beyond the Pocock and O'Brien–Fleming boundaries (1.2) and (1.3). Chapter 6 presents the design details of BHAT and the interim analysis results considered by its Data and Safety Monitoring Board. Inspired by the statistical issues raised by BHAT, a number of important and difficult problems concerning the design and analysis of clinical trials with failure-time endpoints and interim analyses have been resolved in the past 2 decades, and Chap. 6 also describes the "time-sequential" methodology developed in this connection. Chapter 5, however, shows that the fully sequential methodology summarized in Chap. 3 has recently emerged as a standard for prelicensure (Phase III) vaccine safety trials and post-marketing (Phase IV) safety studies.

Analysis of the data at the conclusion of a clinical trial typically involves tests and confidence intervals not only for the primary endpoint but also for different secondary endpoints. The use of a stopping rule whose distribution depends on these parameters introduces substantial difficulties for such inference. Siegmund (1978) developed a method, based on ordering the sample space in a certain way, to construct confidence intervals for its mean of a normal population with known variance following a RST. Alternative orderings of the sample space were subsequently introduced for group sequential tests by Rosner and Tsiatis (1988) and Emerson and Fleming (1990). By making use of resampling methods, Chuang and Lai (1998, 2000) developed a general resampling approach to constructing accurate confidence intervals following sequential tests. Subsequently, Lai and Li (2006) introduced a general ordering scheme that can be used in conjunction with resampling to completely solve the long-standing problem of constructing valid confidence intervals for the primary endpoint of a group sequential trial. Chapter 7 summarizes these developments and describes the methods. Analysis of secondary endpoints following a group sequential trial is also considered in Chap. 7, which reviews the bias-correction approach of Whitehead (1986), Liu et al. (2000), Whitehead et al. (2000), and Hall and Yakir (2003) and describes the hybrid resampling methods of Lai et al. (2009).

1.3 Adaptation and Sequential Optimization

After sequential analysis was introduced in response to more efficient testing of weapons during World War II, it was soon realized that sequential methods could be used to address statistical problems for which there are no solutions with fixed sample sizes. While Dantzig (1940) had shown that no fixed sample size test exists for the problem of testing the null hypothesis $H_0 : \mu = \mu_0$, with prescribed error probabilities α and β at μ_0 and $\mu_0 + \delta$, for the mean μ of a normal distribution whose variance σ^2 is unknown, Stein (1945) showed that a two-stage procedure that uses the first stage to estimate σ^2 and thereby to determine an appropriate second-stage sample size can have power independent of σ. Stein's two-stage design is the first example to show that one can use data during the course of an experiment to learn about the unknown parameters and thereby adapt the experimental design (which is the sample size in Stein's example) as the experiment progresses. It also paved the way for the next generation of adaptive designs in clinical trials in the 1990s that are described in Chap. 8. These adaptive designs, however, are inefficient because they do not incorporate the uncertainties of the parameter estimates at the end of the first stage. Chapter 8 also describes a new class of adaptive designs, introduced by Bartroff and Lai (2008a,b), which use an additional stage to accommodate the uncertainties in the first-stage estimates.

Adaptation via sequential learning of unknown parameters is also a central idea in the theory of nonlinear optimal experimental design. As shown in Sect. 2.3, the optimal design measure involves the unknown parameter vector θ in a nonlinear regression model. To circumvent these difficulties, Fedorov (1972) and others proposed that designs be constructed sequentially, using observations made to date to estimate θ and choosing the next design point by replacing the unknown θ in the optimal design by the current estimate. Lai et al. (2012c) have recently shown the advantages of adaptation in an integrated plan for developing a new drug, which is mentioned earlier in the second paragraph of Sect. 1.1. In the development of a new drug, an important component of the effort and costs involves clinical trials to provide clinical data to support a beneficial claim of the drug and, in case such claim is not valid, to support the termination of the development. The clinical trials progress in steps and are labeled Phase I, II, and III trials, as we have already noted in Sect. 1.1. A project team steers their operations in which intensity, cost, and duration increase with the phase; in particular, Phase III often involves over 3000 professionals, several years to reach completion, and over \$100 million in cost. In addition, there is a core team that makes decisions guided by a clinical development plan (CDP). The CDP maps out the clinical development pathway, beginning with first-in-man studies and ending with submission to the regulatory agency or termination of development. It defines the number and type of clinical studies and their objectives, determines the time sequence of the studies, some of which may be carried out in parallel, identifies key risk areas, and sets key decision points and go/no-go criteria. Julious and Swank (2005) have noted that statistical methods for clinical trial design have focused primarily on “optimizing individual clinical trials” but are lacking “at a more global level in the optimization of clinical development plans.” In practice, however, it is often difficult to specify in advance the cost of each clinical trial in the sequence and the prior probabilities of a go or no-go decision to perform the optimization of CDPs “at a more global level.” Lai et al. (2012c) use ideas from adaptive design of clinical trials, in particular, seamless Phase II–III designs, to adapt a CDP to information acquired during the course of its execution.

Optimization is an important technique in formulating and computing statistical procedures, which can be regarded as statistical decision rules. When the decision rule consists of a sequence of actions, determination of the optimal rule involves dynamic programming. In Chap. 3 we give an introduction to dynamic programming and use it to prove the optimality of the SPRT for simple hypotheses and to derive approximately optimal tests based on generalized likelihood ratio statistics for composite hypotheses. We also give an introduction to recent advances in approximate dynamic programming and apply it to address the treatment versus experimentation dilemma in Phase I cancer trial designs.

1.4 Two Time Scales and Time-Sequential Survival Analysis

As pointed out in Sect. 1.2, the early termination of BHAT paved the way for major advances in the development of group sequential designs. These advances are summarized in Chap. 4, but BHAT and other trials with failure-time endpoints require more subtle methods than those described in Chap. 4. In Chap. 6 we describe these methods that address two time scales in time-sequential survival analysis. “Time-sequential” means that interim analyses are conducted over calendar times, rather than on the time scale measured by the number of subjects at each interim analysis as in group sequential methods in Chap. 4, for which the number of subjects is proportional to the variance (under the null hypothesis) of the test statistic. We begin Chap. 6 with a review of traditional (nonsequential) survival analysis, focusing on how the variances of the commonly used test statistics can be derived with relative ease, despite the complexities due to right censoring, by making use of martingale theory. In the time-sequential setting, calendar time is one time scale, and the other time scale is “information time,” which is measured by the null variance of the test statistic at the time of interim analysis. There is no simple connection between the two time scales and it has been a long-standing problem concerning how to address the difficulties caused by the two time scales in the design and analysis of time-sequential clinical trials with failure-time endpoints.

In Sect. 6.5 we discuss these difficulties and describe the methods that have been developed to address them. These include a comprehensive asymptotic distribution theory for time-sequential censored rank statistics, relatively simple and yet efficient modified Haybittle–Peto tests, and interim Bayesian estimation of the maximum information at the scheduled end of the trial for futility stopping. Section 6.7

describes some recent advances, including the Phase II–III cancer trial designs that we have mentioned in Sect. 1.2 and a method that allows multiple test statistics to increase power if the trial should proceed to its scheduled end. Sections 7.3 and 7.5 describe an innovative hybrid resampling approach to statistical inference from survival data following a time-sequential trial.

1.5 Bayesian and Frequentist Approaches and Associated Software

Berry et al. (2011, p. 1) say that a primary purpose of their book is to describe the Bayesian approach as an alternative to the traditional frequentist approach, which is "the standard statistical approach to designing and analyzing clinical trials and other medical experiments." They find the "flexibility in both design and analysis" and the "decision-oriented" underpinning of the Bayesian approach particularly suited to sequential analysis and adaptive design of clinical trials. On the other hand, they acknowledge that for Phase III confirmatory trials, which are "typically overseen and judged by a regulatory agency," the statistical hurdle for regulatory approval of the new treatment is "to get a statistically significant result at a specified type I error," and the type I error of an adaptive Bayesian design is "extremely difficult, if not impossible, to calculate" and has to be computed by Monte Carlo simulations. Their approach is to adjust the rejection threshold of the Bayesian adaptive/sequential test by using the Monte Carlo simulations carried out under some chosen parameter configuration(s) belonging to the null hypothesis. However, for a composite null hypothesis, there is no guarantee that the worst parameter configuration in the null hypothesis has been chosen for these simulations. An example is given by Lai et al. (2012a, Sect. 4.4) in their comparison of the Phase II–III design described in Chap. 6 with the Bayesian counterpart developed by Huang et al. (2009). Their numerical study shows that because the Bayesian design uses simulations under certain assumed survival rates to control the type I error, the type I error can be substantially inflated under other survival rates belonging to the highly composite null hypothesis. In contrast, the frequentist semiparametric approach used in their design and analysis is shown to maintain the prescribed type I error.

The argument of Berry et al. (2011, Chap. 1) that the Bayesian approach can handle adaptation and sequential learning much more efficiently than the frequentist approach is fair for the prevailing frequentist methods cited in their references, but it overlooks the possibility that *suitably chosen* frequentist methods can work as well, if not better. In fact, there is already a versatile arsenal of statistical methods and theories, including likelihood inference, semiparametric models for censored survival data, bootstrap, and other resampling methods, for nonsequential settings. We shall show in the subsequent chapters how these time-tested methods can be extended to sequential experiments and adaptive designs. In fact, in Chap. 3, we show that these extensions can also be derived as approximations to Bayes rules. Our viewpoint, therefore, is that Bayesian and frequentist approaches should complement each other. One may start with a Bayesian formulation and end up with a frequentist implementation that may be more convenient and appropriate for the problem at hand, for example, confirmatory testing for drug approval. This idea is illustrated in Sect. 3.7 that starts with Bayes sequential tests of one-sided hypotheses and ends up with sequential generalized likelihood ratio tests which have approximately optimal Bayesian and frequentist properties and are also convenient for implementation and description. Another example, which is beyond the scope of this book, is the classical multiarmed bandit problem; see the survey in Lai (2001, pp. 337–339) which shows that while the Bayesian formulation of the infinite-horizon version of the problem has a solution in terms of the "Gittins index" for each arm, a closed-form approximation of the Gittins index yields an upper confidence bound for the arm's mean parameter. Not only does this frequentist approximation provide an intuitive interpretation of the Bayes solution but it also leads to approximately optimal solutions of finite-horizon bandit problems with a frequentist formulation. Conversely, one may start with a frequentist problem and ends up with a Bayes solution. A classic example is the optimality theorem of Wald's SPRT in Sect. 3.6. As explained in Sect. 3.3, Wald conjectured this result on the basis of certain lower bounds for the expected sample sizes under the simple null and alternative hypotheses. Section 3.6 shows that the proof of the conjecture requires solving an auxiliary Bayes problem to which dynamic programming can be applied.

Other than Bayesian designs for Phase I trials, which usually have small sample sizes, considered in Chap. 2 and Sect. 3.8, and the interplay between Bayesian and frequentist approaches to sequential hypothesis testing discussed in Chap. 3, we focus in the subsequent chapters on the frequentist approach and refer readers to the comprehensive treatment of Bayesian methods for clinical trials in Berry et al. (2011). On the other hand, we want to discuss here an irreconcilable difference, which is widely recognized and somewhat controversial, between frequentist and Bayesian inference at the conclusion of a clinical trial with a group sequential or adaptive design. Bayesian inference (e.g., credible sets for parameters) is based on the posterior distribution given the randomly stopped sample, and no adjustment is needed for early stopping or adaptive randomization. In contrast, frequentist inference such as confidence sets has to make adjustments to ensure the correctness of the prescribed coverage probability. In nonsequential designs, the difference between credible and confidence intervals is small for large sample sizes because of the central limit theorem and higher-order expansions of the posterior distribution (Johnson 1970) and the sampling distribution (Gross and Lai 1996) of the approximate pivot used to construct credible or confidence intervals. However, for group sequential designs, this large-sample theory no longer holds, and in fact, an approximate pivot in the fixed sample size case is no longer approximately pivotal in the group sequential setting, as will be explained in Chap. 7 which also describes how valid confidence intervals can be constructed by a resampling procedure, similar to Efron's (1987) bootstrap method to construct confidence intervals based on samples of fixed size.