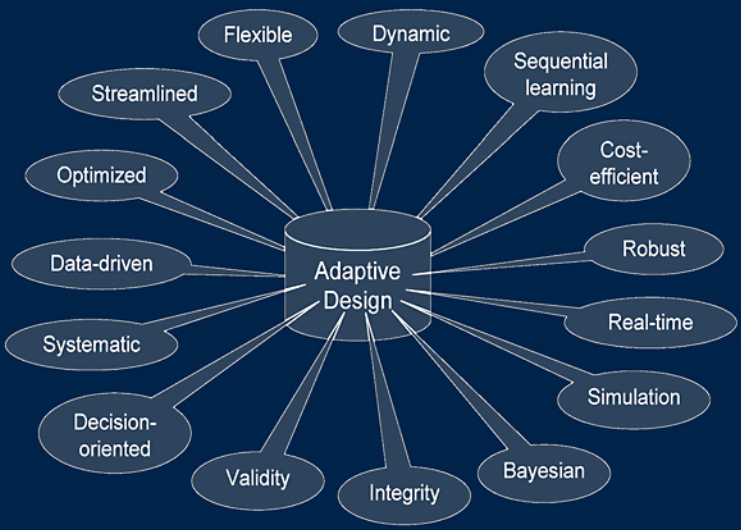
Chapman & Hall/CRC Biostatistics Series

Adaptive Design Theory and Implementation Using SAS and R



Mark Chang

Chapman & Hall/CRC
Taylor & Francis Group

Chapman & Hall/CRC Biostatistics Series

Adaptive Design Theory and Implementation Using SAS and R

Chapman & Hall/CRC Biostatistics Series

Editor-in-Chief

Shein-Chung Chow, Ph.D.
Professor
Department of Biostatistics and Bioinformatics
Duke University School of Medicine
Durham, North Carolina, U.S.A.

Series Editors

Byron Jones
Senior Director
Statistical Research and Consulting Centre
(IPC 193)
Pfizer Global Research and Development
Sandwich, Kent, UK

Jen-pei Liu
Professor
Division of Biometry
Department of Agronomy
National Taiwan University
Taipei, Taiwan

Karl E. Peace
Director, Karl E. Peace Center for Biostatistics
Professor of Biostatistics
Georgia Cancer Coalition Distinguished Cancer Scholar
Georgia Southern University, Statesboro, GA

Chapman & Hall/CRC Biostatistics Series

Adaptive Design Theory and Implementation Using SAS and R

Mark Chang
Millennium Pharmaceuticals
Cambridge, Massachusetts, U.S.A.

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an **informa** business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2008 by Taylor & Francis Group, LLC
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-962-4 (Hardcover)
International Standard Book Number-13: 978-1-58488-962-5 (Hardcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Chang, Mark.
Adaptive design theory and implementation using SAS and R / Mark Chang.
p. ; cm. -- (Chapman & Hall/CRC biostatistics series ; 22)
Includes bibliographical references and index.
ISBN 978-1-58488-962-5 (alk. paper)
1. Clinical trials--Design. 2. Clinical trials--Computer simulation. 3. Clinical trials--Statistical methods. 4. Adaptive sampling (Statistics) 5. SAS (Computer file) 6. R (Computer program language) I. Title. II. Series.
[DNLM: 1. Clinical Trials--methods. 2. Research Design. 3. Biometry--methods. 4. Data Interpretation, Statistical. 5. Software.
QV 20.5 C456a 2008]

R853.C55C42 2008
610.72'4--dc22 2007011412

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To those who are striving toward a better way

.

Series Introduction

The primary objectives of the Biostatistics Book Series are to provide useful reference books for researchers and scientists in academia, industry, and government, and also to offer textbooks for undergraduate and/or graduate courses in the area of biostatistics. This book series will provide comprehensive and unified presentations of statistical designs and analyses of important applications in biostatistics, such as those in biopharmaceuticals. A well-balanced summary will be given of current and recently developed statistical methods and interpretations for both statisticians and researchers/scientists with minimal statistical knowledge who are engaged in the field of applied biostatistics. The series is committed to providing easy-to-understand, state-of-the-art references and textbooks. In each volume, statistical concepts and methodologies will be illustrated through real world examples.

In the past several decades, it is recognized that increasing spending of biomedical research does not reflect an increase of the success rate of pharmaceutical (clinical) development. As a result, the United States Food and Drug Administration (FDA) kicked off a Critical Path Initiative to assist the sponsors in identifying the scientific challenges underlying the medical product pipeline problems. In 2006, the FDA released a Critical Path Opportunities List that outlines 76 initial projects (six broad topic areas) to bridge the gap between the quick pace of new biomedical discoveries and the slower pace at which those discoveries are currently developed into therapies. Among the 76 initial projects, the FDA calls for advancing innovative trial designs, especially for the use of prior experience or accumulated information in trial design. Many researchers interpret it as the encouragement for the use of adaptive design methods in clinical trials.

In clinical trials, it is not uncommon to modify trial and/or statistical procedures during the conduct of clinical trials based on the review of interim data. The purpose is not only to efficiently identify clinical benefits of the test treatment under investigation, but also to increase the probability of success of clinical development. The use of adaptive design methods for modifying the trial and/or statistical procedures of on-going clinical trials based on accrued data has been practiced for years in clinical research. However, it is a concern whether the p-value or confidence interval regarding the treatment effect obtained after the modification is reliable or correct.

In addition, it is also a concern that the use of adaptive design methods in a clinical trial may lead to a totally different trial that is unable to address scientific/medical questions that the trial is intended to answer. In their book, Chow and Chang (2006) provided a comprehensive summarization of statistical methods for the use of adaptive design methods in clinical trials. This volume provides useful approaches for implementation of adaptive design methods in clinical trials through the application of statistical software such as SAS and R. It covers statistical methods for various adaptive designs such as adaptive group sequential design, adaptive dose-escalation design, adaptive seamless phase II/III trial design (drop-the-losers design), and biomarker-adaptive design. It would be beneficial to practitioners such as biostatisticians, clinical scientists, and reviewers in regulatory agencies who are engaged in the areas of pharmaceutical research and development.

Shein-Chung Chow
Editor-in-Chief

Preface

This book is about adaptive clinical trial design and computer implementation. Compared to a classic trial design with static features, an adaptive design allows for changing or modifying the characteristics of a trial based on cumulative information. These modifications are often called adaptations. The word "adaptation" is so familiar to us because we make adaptations constantly in our daily lives according what we learn over time. Some of the adaptations are necessary for survival, while others are made to improve our quality of life. We should be equally smart in conducting clinical trials by making adaptations based on what we learn as a trial progresses. These adaptations are made because they can improve the efficiency of the trial design, provide earlier remedies, and reduce the time and cost of drug development. An adaptive design is also ethically important. It allows for stopping a trial earlier if the risk to subjects outweighs the benefit, or when there is early evidence of efficacy for a safe drug. An adaptive design may allow for randomizing more patients to the superior treatment arms and reducing exposure to inefficacious, but potentially toxic, doses. An adaptive design can also be used to identify better target populations through early biomarker responses.

The aims of this book are to provide a unified and concise presentation of adaptive design theories; furnish the reader with computer programs in SAS and R (also available at www.statisticians.org) for the design and simulation of adaptive trials; and offer (hopefully) a quick way to master the different adaptive designs through examples that are motivated by real issues in clinical trials. The book covers broad ranges of adaptive methods with an emphasis on the relationships among different methods. As Dr. Simon Day pointed out, there are good and bad adaptive designs; a design is not necessarily good just because it is adaptive. There are many rules and issues that must be considered when implementing adaptive designs. This book has included most current regulatory views as well as discussions

of challenges in planning, execution, analysis, and reporting for adaptive designs.

From a "big picture" view, drug development is a sequence of decision processes. To achieve ultimate success, we cannot consider each trial as an isolated piece; instead, a drug's development must be considered an integrated process, using Bayesian decision theory to optimize the design or program as explained in Chapter 16. It is important to point out that every action we take at each stage of drug development is not with the intent of minimizing the number of errors, but minimizing the impact of errors. For this reason, the power of a hypothesis test is not the ultimate criterion for evaluating a design. Instead, many other factors, such as time, safety, and the magnitude of treatment difference, have to be considered in a utility function. From an even bigger-picture view, we are working in a competitive corporate environment, and statistical game theory will provide the ultimate tool for drug development. In the last chapter of the book, I will pursue an extensive discussion of the controversial issues about statistical theories and the fruitful avenues for future research and application of adaptive designs.

Adaptive design creates a new landscape of drug development. The statistical methodology of adaptive design has been greatly advanced by literature in recent years, and there are an increasing number of trials with adaptive features. The PhRMA and BIO adaptive design working groups have made great contributions in promoting innovative approaches to trial design. In preparing the manuscript of this book, I have benefited from discussions with following colleagues: Shein-Chung Chow, Michael Krams, Donald Berry, Jerry Schindler, Michael Chernick, Bruce Turnbull, Barry Turnbull, Sue-Jane Wang (FDA), Vladimir Dragalin, Qing Liu, Simon Day (MHRA), Susan Kenley, Stan Letovsky, Yuan-Yuan Chiu, Jonca Bull, Gorden Lan, Song Yang, Gang Chen, Meiling Lee, Alex Whitmore, Cyrus Mehta, Carl-Fredrik Burman, Richard Simon, George Chi, James Hung (FDA), Aloka Chakravarty (FDA), Marc Walton (FDA), Robert O'Neill (FDA), Paul Gallo, Christopher Jennison, Jun Shao, Keaven Anderson, Martin Posch, Stuart Pocock, Wassmer Gernot, Andy Grieve, Christy Chung, Jeff Maca, Alun Bedding, Robert Hemmings (MHRA), Jose Pinheiro, Jeff Maca, Katherine Sawyer, Sara Radcliffe, Jessica Oldham, Christian Sonesson, Inna Perevozskaya, Anastasia Ivanova, Brenda Gaydos, Frank Bretz, Wenjin Wang, Suman Bhattacharya, and Judith Quinlan.

I would like to thank Hua Liu, PhD, Hugh Xiao, PhD, Andy Boral, MD, MingXiu Hu, PhD, Alun Bedding, PhD, and Jing Xu, PhD, for their careful review and many constructive comments. Thanks to Steve Lewitzky, MS, Kate Rinard, MS, and Frank Chen, MS, Hongliang Shi, MS, Tracy

Zhang, MS, and Rachel Neuwirth MS for support. I wish to express my gratitude to the following individuals for sharing their clinical, scientific, and regulatory insights about clinical trials: Andy Boral, MD, Iain Web, MD, Irvin Fox, MD, Jim Gilbert, MD, Ian Walters, MD, Bill Trepicchio, PhD, Mike Cooper, MD, Dixie-Lee Esseltine, MD, Jing Marantz, MD, Chris Webster, and Robert Pietrusko, Pharm D.

Thanks to Jane Porter, MS, Nancy Simonian, MD, and Lisa Aldler, BA for their support during the preparation of this book. Special thanks to Lori Engelhardt, MA, ELS, for careful reviews and many editorial comments.

From Taylor and Francis, I would like to thank David Grubbs, Sunil Nair, Jay Margolis, and Amber Donley for providing me the opportunity to work on this book.

Mark Chang
Millennium Pharmaceuticals, Inc.,
Cambridge, Massachusetts, USA
Mark.Chang@statisticians.org
www.statisticians.org

Contents

List of Figures

List of Examples

List SAS Macros

List of R Functions

Chapter 1

Introduction

1.1 Motivation

Investment in pharmaceutical research and development has more than doubled in the past decade; however, the increase in spending for biomedical research does not reflect an increased success rate of pharmaceutical development. Figure 1.1 (Data source: PAREXEXL, 2003) illustrates the increase in biomedical research spending and the decrease in NDA (new drug application) submissions over the past ten years.

It is recognized that the increasing spending for biomedical research does not reflect an increased success rate of pharmaceutical development. Reasons for this include (1) a diminished margin for improvement escalates the level of difficulty in proving drug benefits; (2) genomics and other new science have not yet reached their full potential; (3) mergers and other business arrangements have decreased candidates; (4) easy targets are the focus as chronic diseases are more difficult to study; (5) failure rates have not improved; and (6) rapidly escalating costs and complexity decrease willingness/ability to bring many candidates forward into the clinic (Woodcock, 2004).

There are several critical areas for improvement in drug development. One of the obvious areas for improvement is the design, conduct, and analysis of clinical trials. Improvement of the clinical trials process includes (1) the development and utilization of biomarkers or genomic markers, (2) the establishment of quantitative disease models, and (3) the use of more informative designs such as adaptive and/or Bayesian designs. In practice, the use of clinical trial simulation, the improvement of clinical trial monitoring, and the adoption of new technologies for prediction of clinical outcome will also help in increasing the probability of success in the clinical development of promising candidates. Most importantly, we should not use the evaluation tools and infrastructure of the last century to develop this century's

advances. Instead, an innovative approach using adaptive design methods for clinical development must be implemented.

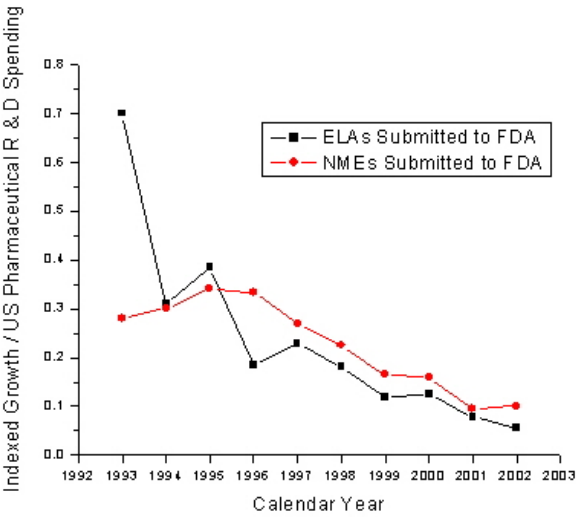


Figure 1.1: Trends in NDAs Submitted to FDA

In the next section, we will provide the definition of adaptive design and brief descriptions of commonly used adaptive designs. In Section 1.3, the importance of computer simulation is discussed. In Section 1.4, we will provide the road map for this book.

1.2 Adaptive Design Methods in Clinical Trials

An adaptive design is a clinical trial design that allows adaptations or modifications to aspects of the trial after its initiation without undermining the validity and integrity of the trial (Chang, 2005; Chow, Chang, and Pong, 2005). The PhRMA Working Group defines an adaptive design as a clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial (Gallo, et al., 2006; Dragalin, 2006).

The adaptations may include, but are not limited to, (1) a group sequential design, (2) an sample-size adjustable design, (3) a drop-losers design, (4) an adaptive treatment allocation design, (5) an adaptive dose-escalation design, (6) a biomarker-adaptive design, (7) an adaptive treatment-switching design, (8) an adaptive dose-finding design, and (9) a combined adaptive

design. An adaptive design usually consists of multiple stages. At each stage, data analyses are conducted, and adaptations are taken based on updated information to maximize the probability of success. An adaptive design is also known as a flexible design (EMEA, 2002).

An adaptive design has to preserve the validity and integrity of the trial. The validity includes internal and external validities. *Internal validity* is the degree to which we are successful in eliminating confounding variables and establishing a cause-effect relationship (treatment effect) within the study itself. A study that readily allows its findings to generalize to the population at large has high *external validity. Integrity* involves minimizing operational bias; creating a scientifically sound protocol design; adhering firmly to the study protocol and standard operating procedures (SOPs); executing the trial consistently over time and across sites or country; providing comprehensive analyses of trial data and unbiased interpretations of the results; and maintaining the confidentiality of the data.

1.2.1 *Group Sequential Design*

A group sequential design (GSD) is an adaptive design that allows for premature termination of a trial due to efficacy or futility, based on the results of interim analyses. GSD was originally developed to obtain clinical benefits under economic constraints. For a trial with a positive result, early stopping ensures that a new drug product can be exploited sooner. If a negative result is indicated, early stopping avoids wasting resources. Sequential methods typically lead to savings in sample-size, time, and cost when compared with the classic design with a fixed sample-size. Interim analyses also enable management to make appropriate decisions regarding the allocation of limited resources for continued development of a promising treatment. GSD is probably one of the most commonly used adaptive designs in clinical trials.

Basically, there are three different types of GSDs: early efficacy stopping design, early futility stopping design, and early efficacy/futility stopping design. If we believe (based on prior knowledge) that the test treatment is very promising, then an early efficacy stopping design should be used. If we are very concerned that the test treatment may not work, an early futility stopping design should be employed. If we are not certain about the magnitude of the effect size, a GSD permitting both early stopping for efficacy and futility should be considered. In practice, if we have a good knowledge regarding the effect size, then a classic design with a fixed sample-size would be more efficient.

1.2.2 *Sample-Size Re-estimation Design*

A sample-size re-estimation (SSR) design refers to an adaptive design that allows for sample-size adjustment or re-estimation based on the review of interim analysis results (Figure 1.2). The sample-size requirement for a trial is sensitive to the treatment effect and its variability. An inaccurate estimation of the effect size and its variability could lead to an underpowered or overpowered design, neither of which is desirable. If a trial is underpowered, it will not be able to detect a clinically meaningful difference, and consequently could prevent a potentially effective drug from being delivered to patients. On the other hand, if a trial is overpowered, it could lead to unnecessary exposure of many patients to a potentially harmful compound when the drug, in fact, is not effective. In practice, it is often difficult to estimate the effect size and variability because of many uncertainties during protocol development. Thus, it is desirable to have the flexibility to re-estimate the sample-size in the middle of the trial.

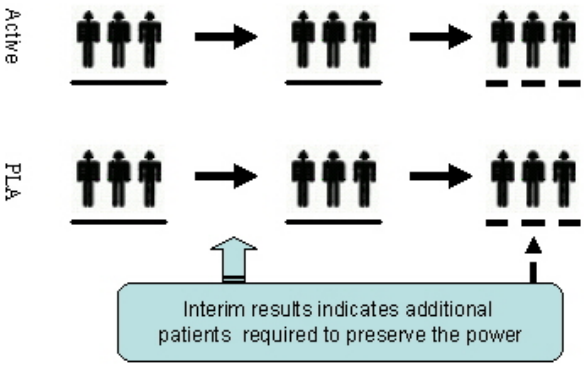


Figure 1.2: Sample-Size Re-Estimation Design

There are two types of sample-size re-estimation procedures, namely, sample-size re-estimation based on blinded data and sample-size re-estimation based on unblinded data. In the first scenario, the sample adjustment is based on the (observed) pooled variance at the interim analysis to recalculate the required sample-size, which does not require unblinding the data. In this scenario, the type-I error adjustment is practically negligible. In the second scenario, the effect size and its variability are re-assessed, and sample-size is adjusted based on the updated information. The statistical method for adjustment could be based on effect size or the conditional power.

Note that the flexibility in SSR is at the expense of a potential loss of power. Therefore, it is suggested that an SSR be used when there are no

good estimates of the effect size and its variability. In the case where there is some knowledge of the effect size and its variability, a classic design would be more efficient.

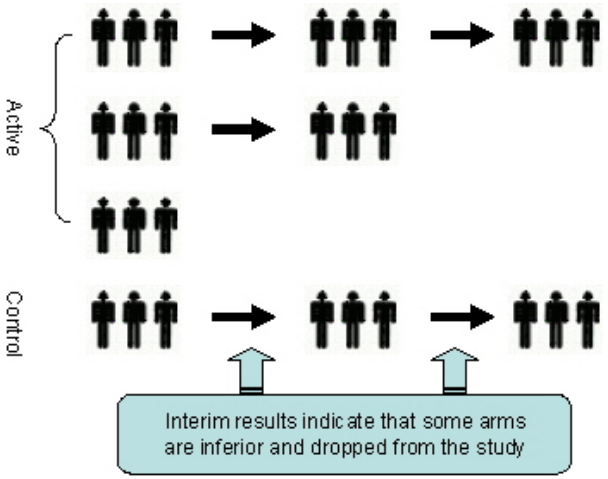


Figure 1.3: Drop-Loser Design

1.2.3 *Drop-Loser Design*

A drop-loser design (DLD) is an adaptive design consisting of multiple stages. At each stage, interim analyses are performed and the losers (i.e., inferior treatment groups) are dropped based on prespecified criteria (Figure 1.3). Ultimately, the best arm(s) are retained. If there is a control group, it is usually retained for the purpose of comparison. This type of design can be used in phase-II/III combined trials. A phase-II clinical trial is often a dose-response study, where the goal is to assess whether there is treatment effect. If there is treatment effect, the goal becomes finding the appropriate dose level (or treatment groups) for the phase-III trials. This type of traditional design is not efficient with respect to time and resources because the phase-II efficacy data are not pooled with data from phase-III trials, which are the pivotal trials for confirming efficacy. Therefore, it is desirable to combine phases II and III so that the data can be used efficiently, and the time required for drug development can be reduced. Bauer and Kieser (1999) provided a two-stage method for this purpose, where investigators can terminate the trial entirely or drop a subset of treatment groups for lack of efficacy after the first stage. As pointed out by Sampson and Sill (2005), the procedure of dropping the losers is highly flexible, and the distributional assumptions are kept to a minimum. However, because of

the generality of the method, it is difficult to construct confidence intervals. Sampson and Sill (2005) derived a uniformly most powerful, conditionally unbiased test for a normal endpoint.

1.2.4 *Adaptive Randomization Design*

An adaptive randomization/allocation design (ARD) is a design that allows modification of randomization schedules during the conduct of the trial. In clinical trials, randomization is commonly used to ensure a balance with respect to patient characteristics among treatment groups. However, there is another type of ARD, called response-adaptive randomization (RAR), in which the allocation probability is based on the response of the previous patients. RAR was initially proposed because of ethical considerations (i.e., to have a larger probability to allocate patients to a superior treatment group); however, response randomization can be considered a drop-loser design with a seamless allocation probability of shifting from an inferior arm to a superior arm. The well-known response-adaptive models include the randomized play-the-winner (RPW) model (see Figure 1.4), an optimal model that minimizes the number of failures. Other response-adaptive randomizations, such as utility-adaptive randomization, also have been proposed, which are combinations of response-adaptive and treatment-adaptive randomization (Chang and Chow, 2005).

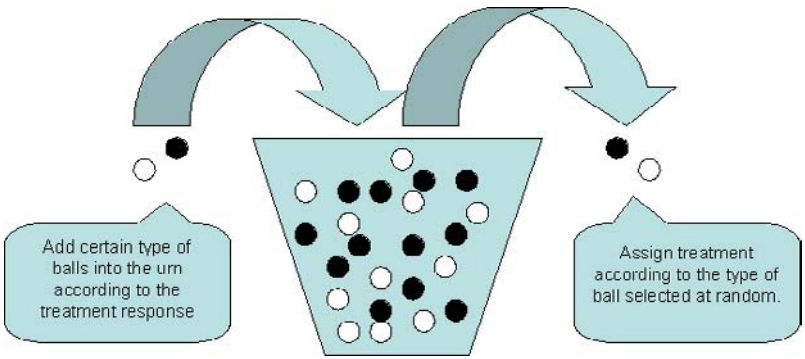


Figure 1.4: Response Adaptive Randomization

1.2.5 *Adaptive Dose-Finding Design*

Dose-escalation is often considered in early phases of clinical development for identifying maximum tolerated dose (MTD), which is often considered the optimal dose for later phases of clinical development. An adaptive dose-finding (or dose-escalation) design is a design at which the dose level used to

treat the next-entered patient is dependent on the toxicity of the previous patients, based on some traditional escalation rules (Figure 1.5). Many early dose-escalation rules are adaptive, but the adaptation algorithm is somewhat ad hoc. Recently more advanced dose-escalation rules have been developed using modeling approaches (frequentist or Bayesian framework) such as the continual reassessment method (CRM) (O'Quigley, et al., 1990; Chang and Chow, 2005) and other accelerated escalation algorithms. These algorithms can reduce the sample-size and overall toxicity in a trial and improve the accuracy and precision of the estimation of the MTD. Note that CRM can be viewed as a special response-adaptive randomization.

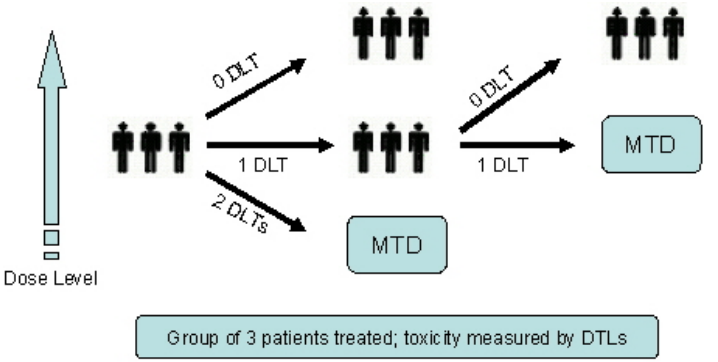


Figure 1.5: Dose-Escalation for Maximum Tolerated Dose

1.2.6 *Biomarker-Adaptive Design*

Biomarker-adaptive design (BAD) refers to a design that allows for adaptations using information obtained from biomarkers. A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic or pathogenic processes or pharmacologic response to a therapeutic intervention (Chakraverty, 2005). A biomarker can be a classifier, prognostic, or predictive marker.

A classifier biomarker is a marker that usually does not change over the course of the study, like DNA markers. Classifier biomarkers can be used to select the most appropriate target population, or even for personalized treatment. Classifier markers can also be used in other situations. For example, it is often the case that a pharmaceutical company has to make a decision whether to target a very selective population for whom the test drug likely works well or to target a broader population for whom the test drug is less likely to work well. However, the size of the selective population may be too small to justify the overall benefit to the patient population. In this case, a BAD may be used, where the biomarker response at in-

terim analysis can be used to determine which target populations should be focused on (Figure 1.6).

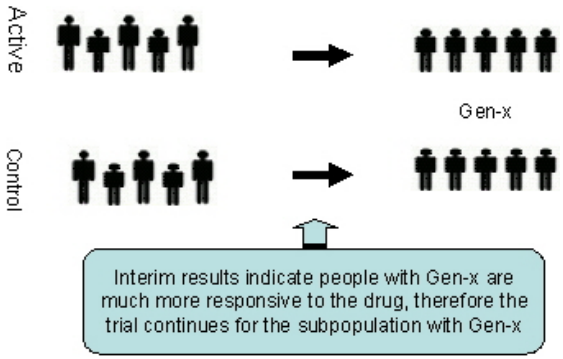


Figure 1.6: Biomarker-Adaptive Design

A prognostic biomarker informs the clinical outcomes, independent of treatment. They provide information about the natural course of the disease in individuals who have or have not received the treatment under study. Prognostic markers can be used to separate good- and poor-prognosis patients at the time of diagnosis. If expression of the marker clearly separates patients with an excellent prognosis from those with a poor prognosis, then the marker can be used to aid the decision about how aggressive the therapy needs to be.

A predictive biomarker informs the treatment effect on the clinical endpoint. Compared to a gold-standard endpoint, such as survival, a biomarker can often be measured earlier, easier, and more frequently. A biomarker is less subject to competing risks and less affected by other treatment modalities, which may reduce sample-size due to a larger effect size. A biomarker could lead to faster decision-making. However, validating predictive biomarkers is challenging. BAD simplifies this challenge. In a BAD, "softly" validated biomarkers are used at the interim analysis to assist in decision-making, while the final decision can still be based on a gold-standard endpoint, such as survival, to preserve the type-I error (Chang, 2005).

1.2.7 *Adaptive Treatment-Switching Design*

An adaptive treatment-switching design (ATSD) is a design that allows the investigator to switch a patient's treatment from the initial assignment if there is evidence of lack of efficacy or a safety concern (Figure 1.7).

To evaluate the efficacy and safety of a test treatment for progressive

diseases, such as cancers and HIV, a parallel-group, active-control, randomized clinical trial is often conducted. In this type of trial, qualified patients are randomly assigned to receive either an active control (a standard therapy or a treatment currently available in the marketplace) or a test treatment under investigation. Due to ethical considerations, patients are allowed to switch from one treatment to another if there is evidence of lack of efficacy or disease progression. In practice, it is not uncommon that up to 80% of patients may switch from one treatment to another. Sommer and Zeger (1991) referred to the treatment effect among patients who complied with treatment as "biological efficacy." Branson and Whitehead (2002) widened the concept of biological efficacy to encompass the treatment effect as if all patients adhered to their original randomized treatments in clinical studies allowing treatment switching. Despite allowing a switch in treatment, many clinical studies are designed to compare the test treatment with the active control agent as if no patients had ever been switched. This certainly has an impact on the evaluation of the efficacy of the test treatment, because the response-informative switching causes the treatment effect to be confounded. The power for the methods without considering the switching is often lost dramatically because many patients from two groups eventually took the same drugs (Shao, Chang, and Chow, 2005). Currently, more approaches have been proposed, which include mixed exponential mode (Chang, 2006, Chow, and Chang, 2006) and a mixture of the Wiener processes (Lee, Chang, and Whitmore, 2007).

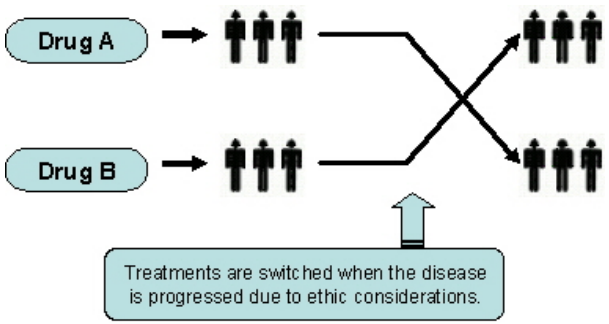


Figure 1.7: Adaptive Treatment Switching

1.2.8 *Clinical Trial Simulation*

Clinical trial simulation (CTS) is a process that mimics clinical trials using computer programs. CTS is particularly important in adaptive designs for several reasons: (1) the statistical theory of adaptive design is complicated with limited analytical solutions available under certain assumptions; (2)

the concept of CTS is very intuitive and easy to implement; (3) CTS can be used to model very complicated situations with minimum assumptions, and type-I error can be strongly controlled; (4) using CTS, we can not only calculate the power of an adaptive design, but we can also generate many other important operating characteristics such as expected sample-size, conditional power, and repeated confidence interval - ultimately this leads to the selection of an optimal trial design or clinical development plan; (5) CTS can be used to study the validity and robustness of an adaptive design in different hypothetical clinical settings, or with protocol deviations; (6) CTS can be used to monitor trials, project outcomes, anticipate problems, and suggest remedies before it is too late; (7) CTS can be used to visualize the dynamic trial process from patient recruitment, drug distribution, treatment administration, and pharmacokinetic processes to biomarkers and clinical responses; and finally, (8) CTS has minimal cost associated with it and can be done in a short time.

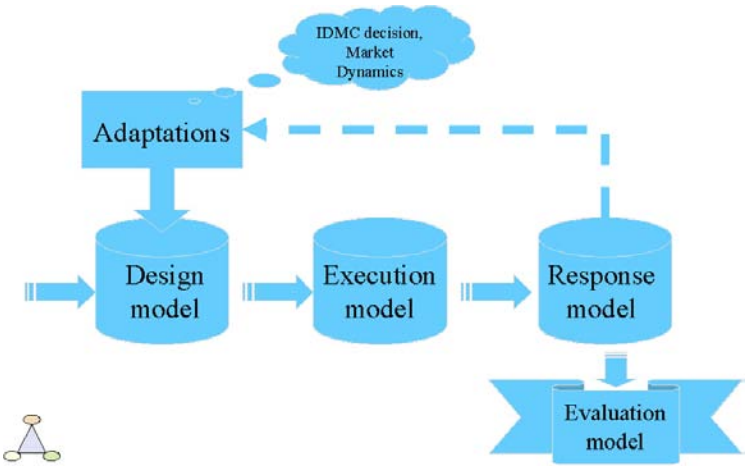


Figure 1.8: Clinical Trial Simulation Model

CTS was started in the early 1970s and became popular in the mid 1990s due to increased computing power. CTS components include (1) a Trial Design Mode, which includes design type (parallel, crossover, traditional, adaptive), dosing regimens or algorithms, subject selection criteria, and time, financial, and other constraints; (2) a Response Model, which includes disease models that imitate the drug behavior (PK and PD models) or intervention mechanism, and an infrastructure model (e.g., timing and validity of the assessment, diagnosis tool); (3) an Execution Model, which models the human behaviors that affect trial execution (e.g., protocol compliance, cooperation culture, decision cycle, regulatory authority, inference

of opinion leaders); and (4) an Evaluation Model, which includes criteria for evaluating design models, such as utility models and Bayesian decision theory. The CTS model is illustrated in Figure 1.8.

1.2.9 *Regulatory Aspects*

The FDA's Critical Path initiative is a serious attempt to bring attention and focus to the need for targeted scientific efforts to modernize the techniques and methods used to evaluate the safety, efficacy, and quality of medical products as they move from product selection and design to mass manufacture. Critical Path is NOT about the drug discovery process. The FDA recognizes that improvement and new technology are needed. The National Institutes of Health (NIH) is getting more involved via the "roadmap" initiative. Critical Path is concerned with the work needed to move a candidate all the way to a marketed product. It is clear that the FDA supports and encourages innovative approaches in drug development. The regulatory agents feel that some adaptive designs are encouraging, but are cautious about others, specially for pivotal studies (Temple, 2006; Hung, et el., 2006; Hung, Wang, and O'Neill, 2006; EMEA, 2006).

In the past five years FDA has received different adaptive design protocols. The design adaptations FDA reviewers have encountered are: extension of sample-size, termination of a treatment arm, change of the primary endpoint, change of statistical tests, and change of the study objective such as from superiority to non-inferiority or vice versa, and selection of a subgroup based upon externally available studies (Hung, O'Neill, Wang, and Lawrence, 2006). Dr. O'Neill from FDA shared two primary motivations that may explain why adaptive or flexible designs might be useful. One is the goal of an adaptive/flexible design to allow some type of mid-study changes that are prospectively planned in order to maximize the chance of success of the trial while properly preserving the type-I error rate because some planning parameters are imprecisely known. Another goal is to enrich trials with subgroups of patients having genomic profiles likely to respond or less likely to experience toxicity (Hung, O'Neill, Wang, and Lawrence, 2006).

"Adaptive designs should be encouraged for Phases I and II trials for better exploration of drug effects, whether beneficial or harmful, so that such information can be more optimally used in latter stages of drug development. Controlling false positive conclusions in exploratory phases is also important so that the confirmatory trials in latter stages achieve their goals. The guidance from such trials properly controlling false positives may be more informative to help better design confirmatory trials." (Hung,

O'Neill, Wang, and Lawrence, 2006). As pointed out by FDA statistician Dr. Stella Machado, "The two major causes of delayed approval and nonapproval of phase III studies is poor dose selection in early studies and phase III designs [that] don't utilize information from early phase studies" ("The Pink Sheet", Dec. 18, 2006, p.24). FDA is granting industry a great deal of leeway in adaptive design in early learning phase, at the same time suggests that emphasis be placed on dose-response and exposure risk. Dr. O'Neill said that learning about the dose-response relationship lies at the heart of adaptive designs ("The Pink Sheet", Dec. 18, 2006, p.24). Companies should begin a dialogue about adaptive designs with FDA medical officers and statisticians as early as a year before beginning a trial as suggested by Dr.Robert Powell from FDA ("The Pink Sheet", Dec. 18, 2006, p.24).

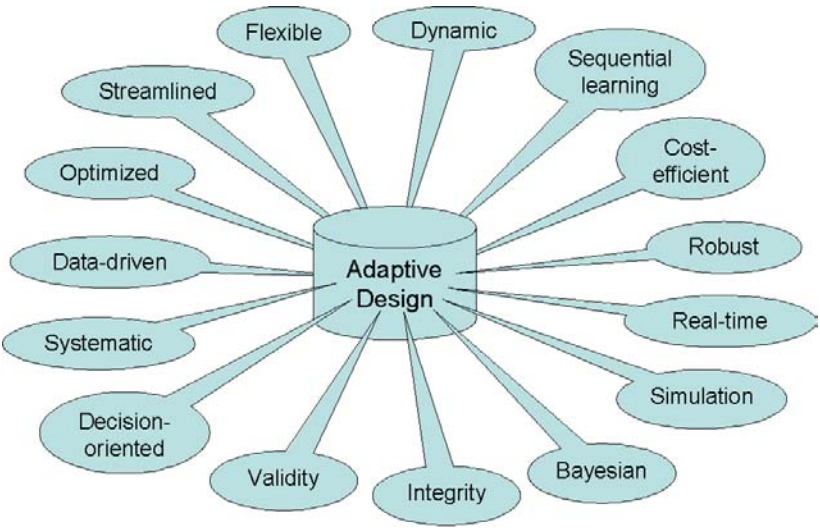


Figure 1.9: Characteristics of Adaptive Designs

1.2.10 *Characteristics of Adaptive Designs*

Adaptive design is a sequential data-driven approach. It is a dynamic process that allows for real-time learning. It is flexible and allows for modifications to the trial, which make the design cost-efficient and robust against the failure. Adaptive design is a systematic way to design different phases of trials, thus streamlining and optimizing the drug development process. In contrast, the traditional approach is composed of weakly connected phase-wise processes. Adaptive design is a decision-oriented, sequential learning process that requires up-front planning and a great deal of collaboration among the different parties involved in the drug development process. To

this end, Bayesian methodology and computer simulation play important roles. Finally, the flexibility of adaptive design does not compromise the validity and integrity of the trial or the development process (Figure 1.9).

Adaptive design methods represent a revolution in pharmaceutical research and development. Using adaptive designs, we can increase the chances for success of a trial with a reduced cost. Bayesian approaches provide an ideal tool for optimizing trial designs and development plans. Clinical trial simulations offer a powerful tool to design and monitor trials. Adaptive design, the Bayesian approach, and trial simulation combine to form an ultimate statistical instrument for the most successful drug development programs.

1.3 FAQs about Adaptive Designs

In recent years, I was interviewed by several journalists from scientific and technological journals including Nature Biotechnology, BioIT World, and Contract Pharms among others. The following are some questions that were commonly asked about adaptive designs.

1. *What is the classification of an adaptive clinical trial? Is there a consensus in the industry regarding what adaptive trials entail?*

After many conferences and discussions, there is more or less a consensus on the definition of adaptive design. A typical definition is as follows:

An adaptive design is a design that allows modifications to aspects of the trial after its initiation without undermining the validity and integrity of the trial. All adaptive designs involve interim analyses and adaptations or decision-making based on the interim results.

There are many ways to classify adaptive designs. The following are the common examples of adaptive trials:

- Sample size re-estimation design to increase the probability of success
- Early stopping due to efficacy or futility design to reduce cost and time
- Response adaptive randomization design to give patients a better chance of assigning to superior treatment
- Drop-loser design for adaptive dose finding to reduce sample-size by dropping the inferior treatments earlier
- Adaptive dose escalation design to minimize toxicity while at the same time acquiring information on maximum tolerated dose
- Adaptive seamless design combining two traditional trials in different phases into a single trial, reducing cost and time to market
- Biomarker-adaptive design to have earlier efficacy or safety readout

to select better target populations or subpopulation

2. *What challenges does the adaptive trial model present?*

Adaptive designs can reduce time and cost, minimize toxicity, and help select the best dose for the patients and better target populations. With adaptive design, we can develop better science for testing new drugs, and in turn, better science for prescribing them.

There are challenges associated with adaptive design. Statistical methods are available for most common adaptive designs, but for more complicated adaptive designs, the methodologies are still in development.

Operationally, an adaptive design often requires real-time or near real-time data collection and analysis. In this regard, data standardizations, such as CDISC and electronic data capture (EDC), are very helpful in data cleaning and reconciliation. Note that not all adaptive designs require perfectly clean data at interim analysis, but the cleaner the data are, the more efficient the design is. Adaptive designs require the ability to rapidly integrate knowledge and experiences from different disciplines into the decision-making process and hence require a shift to a more collaborative working environment among disciplines.

From a regulatory standpoint, there is no regulatory guidance for adaptive designs at the moment. Adaptive trials are reviewed on a case-by-case basis. Naturally there are fears that a protocol using this innovative approach may be rejected, causing a delay.

The interim unblinding may potentially cause bias and put the integrity of the trial at risk. Therefore, the unblinding procedure should be well established before the trial starts, and frequent unblinding should be avoided. Also, unblinding the premature results to the public could jeopardize the trial.

3. *How would adaptive trials affect traditional phases of drug development? How are safety and efficacy measured in this type of trial?*

Adaptive designs change the way we conduct clinical trials. Trials in different phases can be combined to create a seamless study. The final safety and efficacy requirements are not reduced because of adaptive designs. In fact, with adaptive designs, the efficacy and safety signals are collected and reviewed earlier and more often than in traditional designs. Therefore, we have a better chance of avoiding unsafe drug exposure to large patient populations. A phase-II and III combined seamless design, when the trial is carried out to the final stage, has longer-term patient efficacy and safety data than traditional phase-II, phase-III trials; however, precautions should be taken at the interim decision-making when data are not mature.

4. *If adaptive trials become widely adopted, how would it impact clinical trial materials and the companies that provide them?*

Depending on the type of adaptive design, there might be requirements for packaging and shipping to be faster and more flexible. Quick and accurate efficacy and safety readouts may also be required. The electronic drug packages with an advanced built-in recording system will be helpful.

If adaptive trials become widely adopted, the drug manufacturers who can provide the materials adaptively will have a better chance of success.

5. *What are some differences between adaptive trials and the traditional trial model with respect to the supply of clinical trial materials?*

For a classic design, the amount of material required is fixed and can be easily planned before the trial starts. However, for some adaptive trials, the exact amount of required materials is not clear until later stages of the trial. Also the next dosage for a site may not be fully determined until the time of randomization; therefore, vendors may need to develop a better drug distribution strategy.

6. *What areas of clinical development would experience cost/time savings with the adaptive trial model?*

Adaptive design can be used in any phase, even in the preclinical and discovery phases. Drug discovery and development is a sequence of decision processes. The traditional paradigm breaks this into weakly connected fragments or phases. An adaptive approach will eventually be utilized for the whole development process to get the right drug to the right patient at the right time.

Adaptive design requires fewer patients, less trial material, sometimes fewer lab tests, less work for data collection and fewer data queries to be resolved. However, an adaptive trial requires much more time during upfront planning and simulation studies.

7. *What are some of the regulatory issues that need to be addressed for this type of trial?*

So far FDA is very positive about innovative adaptive designs. Guidance is expected in the near future (see DFA Deputy Commissioner Dr. Scott Gottlieb's speech delivered at the Adaptive Design conference in July 2006).

If the adaptive design is submitted with solid scientific support and strong ethical considerations and it is operationally feasible, there should not be any fears of rejection of such a design. On the other hand, with a significant increase in adaptive trials in NDA submissions, regulatory bodies may face a temporary shortage of resources for reviewing such designs. Adaptive designs are relatively new to the industry and to regulatory bodies; therefore, there is a lot to learn by doing them. For this reason, it is a good idea to start with adaptive designs in earlier stages of drug development.

1.4 Roadmap

Chapter 2, Classic Design: This chapter will review the classic design and issues raised from the traditional approaches. The statistical design methods discussed include one- and two-group designs, multiple-group dose-response designs, as well as equivalence and noninferiority designs.

Chapter 3, Theory of Adaptive Design: This chapter introduces unified theory for adaptive designs, which covers four key statistical elements in adaptive designs: stopping boundary, adjusted p-value, point estimation, and confidence interval. We will discuss how different approaches can be developed under this unified theory and what the common adaptations are.

Chapter 4, Method with Direct Combination of P-values: Using the unified formulation discussed in Chapter 3, the method with an individual stagewise p-value and the methods with the sum and product of the stagewise p-values are discussed in detail for two-stage adaptive designs. Trial examples and step-by-step instructions are provided.

Chapter 5, Method with Inverse-Normal P-values: The Inverse-Normal method generalizes the classic group sequential method. The method can also be viewed as weighted stagewise statistics and includes several other methods as special cases. Mathematical formulations are derived and examples are provided regarding how to use the method for designing a trial.

Chapter 6, Implementation of K-Stage Design: Chapters 4 and 5 are mainly focused on two-stage adaptive designs because these designs are simple and usually have a closed-form solution. In Chapter 6, we use simulation approaches to generalize the methods in Chapters 4 and 5 to K-stage designs using SAS macros and R functions; many examples are provided.

Chapter 7, Conditional Error Function Method: The conditional error function method is a very general approach. We will discuss in particular the Proschan-Hunsberger method and the Muller-Schafer method. We will compare the conditional error functions for various other methods and study the relationships between different adaptive design methods through the conditional error functions and conditional power.

Chapter 8, Recursive Adaptive Design: The recursive two-stage adaptive design not only offers a closed-form solution for K-stage designs, but also allows for very broad adaptations. We first introduce two powerful principles, the error-spending principle and the conditional error principle, from which we further derive the recursive approach. Examples are provided to illustrate the different applications of this method.

Chapter 9, Sample-Size Re-Estimation Design: This chapter is devoted to the commonly used adaptation, sample-size re-estimation. Various sample-size re-estimation methods are evaluated and compared. The

goal is to demonstrate a way to evaluate different methods under different conditions and to optimize the trial design that fits a particular situation. Practical issues and concerns are also addressed.

Chapter 10, Multiple-Endpoint Adaptive Design: One of the most challenging issues is the multiple-endpoint analysis with adaptive design. This is motivated by an actual adaptive design in an oncology trial. The statistical method is developed for analyzing the multiple-endpoint issues for both coprimary and primary-secondary endpoints.

Chapter 11, Drop-Loser and Add-Arm Designs: Drop-loser and add-arm design can be used in adaptive dose-finding studies and combined phase-II and III studies (seamless design). Different drop-loser/add-arm designs are also shown with weak and strong alpha-control in the examples.

Chapter 12, Biomarker Adaptive Design: In this chapter, adaptive design methods are developed for classifier, diagnosis, and predictive markers. SAS macros have been developed for biomarker-adaptive designs. The improvement in efficiency is assessed for difference methods in different scenarios.

Chapter 13, Response-Adaptive Treatment Switching and Crossover: Response-adaptive treatment switching and crossover are statistically challenging. Treatment switching is not required for the statistical efficacy of a trial design; rather, it is motivated by an ethical consideration. Several methods are discussed, including the time-dependent exponential, mixed exponential, and a mixture of Wiener models.

Chapter 14, Response-Adaptive Allocation Design: Response-adaptive randomizations/allocations have many different applications. They can be used to reduce the overall sample-size and the number of patients exposed to ineffective or even toxic regimens. We will discuss some commonly used adaptive randomizations, such as randomized-play-the-winner. Use of response-adaptive randomization for general adaptations is also discussed.

Chapter 15, Adaptive Dose Finding Design: The adaptive dose finding designs, or dose-escalation designs, are discussed in this chapter. The goal is to reduce the overall sample-size and the number of patients exposed to ineffective or even toxic regimens, and to increase the precision and accuracy of MTD (maximum tolerated dose) assessment. We will discuss oncology dose-escalation trials with traditional and Bayesian continual reassessment methods

Chapter 16, Bayesian Adaptive Design: The philosophical differences between the Bayesian and frequentist approaches are discussed. Through many examples, the two approaches are compared in terms of design, monitoring, analysis, and interpretation of results. More importantly, how to use Bayesian decision theory to further improve the efficiency of adaptive

designs is discussed with examples.

Chapter 17, Planning, Execution, Analysis, and Reporting: In this chapter, we discuss the logistic issues with adaptive designs. The topics cover planning, monitoring, analysis, and reporting for adaptive trials. It also includes most concurrent regulatory views and recommendations.

Chapter 18, Debate and Perspectives: This chapter is a future perspective of adaptive designs. We will present very broad discussions of the challenges and controversial presented by adaptive designs from philosophical and statistical perspectives.

Appendix A, Random number generation

Appendix B, R programs for adaptive designs

Computer Programs

Most adaptive design methods have been implemented and tested in SAS 8.0 and 9.0 and major methods have also been implemented in R. These computer programs are compact (often fewer than 50 lines of SAS code) and ready to use. For convenience, electronic versions of the programs have been made available at www.statisticians.org.

The SAS code is enclosed in ››**SAS Macro x.x**›› and ‹‹**SAS**‹‹ or in ››**SAS**›› and ‹‹**SAS**‹‹. R programs are presented in Appendix B.

Chapter 2

Classic Design

2.1 Overview of Drug Development

Pharmaceutical medicine uses all the scientific, clinical, statistical, regulatory, and business knowledge available to provide a challenging and rewarding career. On average, it costs about $1.8 billion to take a new compound to market and only one in 10,000 compounds ever reach the market. There are three major phases of drug development: (1) preclinical research and development, (2) clinical research and development, and (3) after the compound is on the market, a possible "post-marketing" phase

The preclinical phase represents bench work (in vitro) followed by animal testing, including kinetics, toxicity, and carcinogenicity. An investigational new drug application (IND) is submitted to the FDA seeking permission to begin the heavily regulated process of clinical testing in human subjects. The clinical research and development phase, representing the time from the beginning of human trials to the new drug application (NDA) submission that seeks permission to market the drug, is by far the longest portion of the drug development cycle and can last from 2 to 10 years (Tonkens, 2005).

Clinical trials are usually divided into three phases. The primary objectives of phase I are to (1) determine the metabolism and pharmacological activities of the drug, the side effects associated with increasing dose, and early evidence of effectiveness, and (2) to obtain sufficient information regarding the drug's pharmacokinetics and pharmacological effects to permit the design of well-controlled and scientifically valid phase-II clinical studies (21 CFR 312.21). Unless it is an oncology study, where the maximum tolerated dose (MTD) is primarily determined by a phase-I dose-escalation study, the dose-response or dose-finding study is usually conducted in phase II, and efficacy is usually the main focus. The choice of study design and study population in a dose-response trial will depend on the phase of de-

velopment, therapeutic indication under investigation, and severity of the disease in the patient population of interest (ICH Guideline E4, 1994). Phase-III trials are considered confirmative trials.

The FDA does not actually approve the drug itself for sale. It approves the labeling, the package insert. United States law requires truth in labeling, and the FDA ensures that claims that a drug is safe and effective for treatment of a specified disease or condition have, in fact, been proven. All prescription drugs must have labels, and without proof of the truth of its label, a drug may not be sold in the United States.

In addition to mandated conditional regulatory approval and post-marketing surveillance trials, other reasons sponsors may conduct post-marketing trials include comparing their drug with that of competitors, widening the patient population, changing the formulation or dose regimen, or applying a label extension. A simplified view of the NDA is shown in Figure 2.1 (Tonkens, 2005).

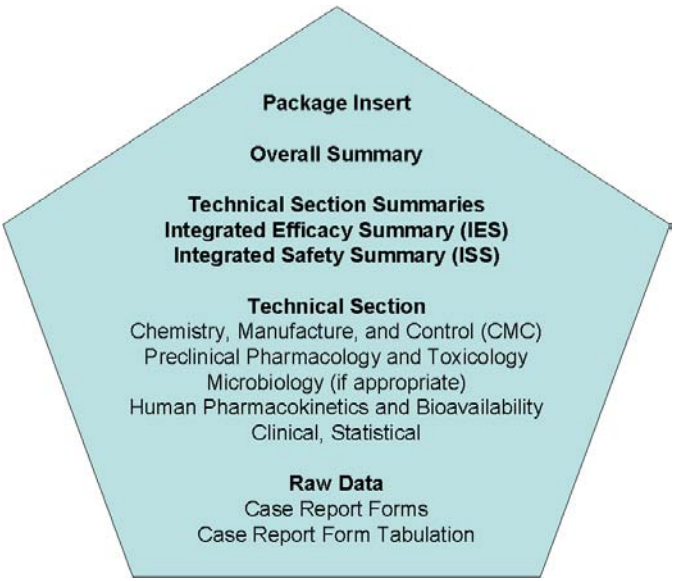


Figure 2.1: A Simplified View of the NDA

In classic trial designs, power and sample-size calculations are a major task. The sample-size calculations for two-group designs have been studied by many scholars, among them Julious (2004), Chow, et al.(2003), Machin, et al. (1997), Campbell, et al. (1995), and Lachin and Foukes (1986).

In what follows, we will review a unified formulation for sample-size calculation in classic two-arm designs including superiority, noninferiority,

and equivalence trials. We will also discuss some important concepts and issues with the designs that are often misunderstood. We will first discuss two-group superiority and noninferiority designs in Section 2.2. Equivalence studies will be discussed in Section 2.3. Three different types of equivalence studies (average, population, and individual equivalences) are reviewed. We will discuss dose-response studies in Section 2.4. The sample-size calculations for various endpoints are provided based on the contrast test. Section 2.5 will discuss the maximum information design, in which the sample-size changes automatically according the variance.

2.2 Two-Group Superiority and Noninferiority Designs

2.2.1 *General Approach to Power Calculation*

When testing a null hypothesis $H_o : \varepsilon \leq 0$ against an alternative hypothesis $H_a : \varepsilon > 0$, where ε is the treatment effect (difference in response), the type-I error rate function is defined as

$$\alpha(\varepsilon) = \Pr\{\text{reject } H_o \text{ when } H_o \text{ is true}\}.$$

Note: alternatively, the type-I error rate can be defined as $\sup_{\varepsilon \in H_o}\{\alpha(\varepsilon)\}$. Similarly, the type-II error rate function β is defined as

$$\beta(\varepsilon) = \Pr\{\text{fail to reject } H_o \text{ when } H_a \text{ is true}\}.$$

For hypothesis testing, knowledge of the distribution of the test statistic under H_o is required. For sample-size calculation, knowledge of the distribution of the test statistic under a particular H_a is also required. To control the overall type-I error rate at level α under any point of the H_o domain, the condition $\alpha(\varepsilon) \leq \alpha^*$ for all $\varepsilon \leq 0$ must be satisfied, where α^* is a threshold that is usually larger than 0.025 unless it is a phase III trial. If $\alpha(\varepsilon)$ is a monotonic function of ε, then the maximum type-I error rate occurs when $\varepsilon = 0$, and the test statistic should be derived under this condition. For example, for the null hypothesis $H_o : \mu_2 - \mu_1 \leq 0$, where μ_1 and μ_2 are the means of the two treatment groups, the maximum type-I error rate occurs on the boundary of H_o when $\mu_2 - \mu_1 = 0$. Let $T = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}}$, where $\hat{\mu}_i$ and $\hat{\sigma}$ are the sample mean and pooled sample standard deviation, respectively. Further, let $\Phi_o(T)$ denote the cumulative distribution function (c.d.f) of the test statistic on the boundary of the null hypothesis domain, and let $\Phi_a(T)$ denote the c.d.f under H_a. Given this information, under the large sample

assumption, $\Phi_o(T)$ is the c.d.f of the standard normal distribution, $N(0,1)$, and $\Phi_a(T)$ is the c.d.f. of $N(\frac{\sqrt{n}\varepsilon}{2\sigma},1)$, where n is the total sample-size and σ is the common standard deviation (Figure 2.2).

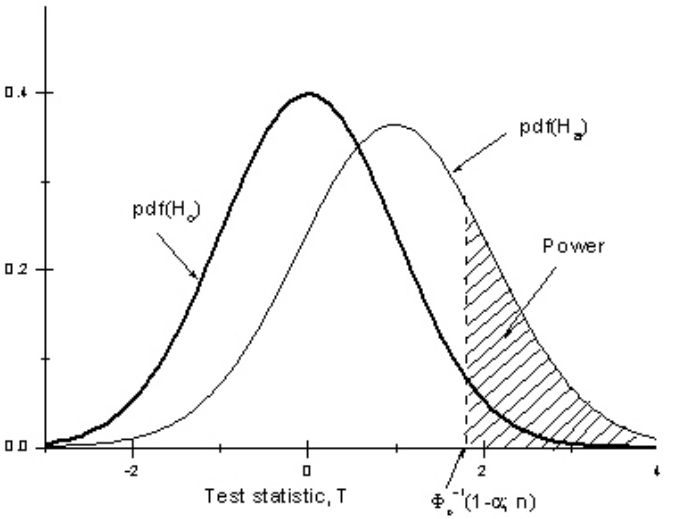


Figure 2.2: Power as a Function of α and n

The power of the test statistic T under a particular H_a can be expressed as follows:

$$\text{Power}\,(\varepsilon) = \Pr(T \geq \Phi_o^{-1}(1-\alpha;n)|H_a) = 1 - \Phi_a(\Phi_o^{-1}(1-\alpha;n);n),$$

which is equivalent to

$$\text{Power}\,(\varepsilon) = \Phi\left(\frac{\sqrt{n}\varepsilon}{2\sigma} - z_{1-\alpha}\right), \tag{2.1}$$

where Φ is the c.d.f of the standard normal distribution, ε is treatment difference, and $z_{1-\beta}$ and $z_{1-\alpha}$ are the percentiles of the standard normal distribution. Figure 2.2 is an illustration of the power function of α and the sample-size n. The total sample-size is given by

$$n = \frac{4(z_{1-a} + z_{1-\beta})^2 \sigma^2}{\varepsilon^2}. \tag{2.2}$$

More generally, for an imbalanced design with sample-size ratio $r = n_1/n_2$ and a margin δ ($\delta > 0$ for superiority test and $\delta < 0$ for non-inferiority test), the sample-size is given by

$$n_2 = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \sigma^2 \left(1 + 1/r\right)}{\left(\varepsilon - \delta\right)^2}. \tag{2.3}$$

(2.3) is a general sample-size formulation for the two-group designs with a normal, binary, or survival endpoint. When using the formulation, the corresponding "standard deviation" σ should be used, examples of which have been listed in Table 2.1 for commonly used endpoints (Chang and Chow, 2006a).

We now derive the standard deviation for the time-to-event endpoint. Under an exponential survival model, the relationship between hazard (λ), median (T_{median}) and mean (T_{mean}) survival time is very simple:

$$T_{Median} = \frac{\ln 2}{\lambda} = (\ln 2) T_{mean}.$$

Let λ_i be the population hazard rate for group i. The corresponding variance σ_i^2 can be derived in several different ways. Here we use Lachin and Foulkes' maximum likelihood approach (Lachin and Foulkes 1986 and Chow, Shao, and Wang 2003).

Let T_0 and T_s be the accrual time period and the total trial duration, respectively. We then can prove that the variance for uniform patient entry is given by

$$\sigma^2(\lambda_i) = \lambda_i^2 \left[1 + \frac{e^{-\lambda_i T_s}(1 - e^{\lambda_i T_0})}{T_0 \lambda_i}\right]^{-1}.$$

Let a_{ij} denote the uniform entry time of the j^{th} patient of the i^{th} group, i.e., $a_{ij} \sim \frac{1}{T_0}$, $0 \le a_{ij} \le T_0$. Let t_{ij} be the time-to-event starting from the time of the patient's entry for the j^{th} patient in the i^{th} group, $i = 1, ...k,\ j = 1, ..., n_i$. It is assumed that t_{ij} follows an exponential distribution with a hazard rate of λ_i. The information observed is (x_{ij}, δ_{ij}) $= (\min(t_{ij},\ T_s - a_{ij}),\ I\{t_{ij} \le T_s - a_{ij}\})$. For a fixed i, the joint likelihood for $x_{ij},\ j = 1, ...n_i$ can be written as

$$L(\lambda_i) = \frac{1}{T_0}\lambda_i^{\sum_{j=1}^{n_i}\delta_{ij}} e^{-\lambda_i \sum_{j=1}^{n_i} x_{ij}}.$$

Taking the derivative with respect to λ_i and letting it equal zero, we can obtain the maximum likelihood estimate (MLE) for λ_i, which is given by $\hat{\lambda}_i = \frac{\sum_{j=1}^{n_i}\delta_{ij}}{\sum_{j=1}^{n_i} x_{ij}}$. According to the Central Limit Theorem, we have

$$\begin{aligned}\sqrt{n_i}(\hat{\lambda}_i - \lambda_i) &= \sqrt{n_i}\frac{\sum_{j=1}^{n_i}(\delta_{ij} - \lambda_i x_{ij})}{\sum_{j=1}^{n_i} x_{ij}} \\ &= \frac{1}{\sqrt{n_i}E(x_{ij})}\sum_{j=1}^{n_i}(\delta_{ij} - \lambda_i x_{ij}) + o_p(1) \\ &\xrightarrow{d} N(0, \sigma^2(\lambda_i)),\end{aligned}$$

where

$$\sigma^2(\lambda_i) = \frac{var(\delta_{ij} - \lambda_i x_{ij})}{E^2(x_{ij})}$$

and $\xrightarrow{d}$ denotes convergence in distribution. Note that

$$E(\delta_{ij}) = E(\delta_{ij}^2) = 1 - \int_0^{T_0}\frac{1}{T_0}e^{-\lambda_i(T_s - a)}da = 1 + \frac{e^{-\lambda_i T_s}(1 - e^{\lambda_i T_0})}{T_0\lambda_i}$$

$$E(x_{ij}) = \frac{1}{\lambda_i}E(\delta_{ij}), \text{ and } E(x_{ij}^2) = \frac{2E(\delta_{ij}x_{ij})}{\lambda_i}.$$

Hence,

$$\begin{aligned}\sigma^2(\lambda_i) &= \frac{var(\delta_{ij} - \lambda_i x_{ij})}{E^2(x_{ij})} = \frac{1}{E^2(x_{ij})}\left(E(\delta_{ij}^2) - 2\lambda_i E(\delta_{ij}x_{ij}) + \lambda_i^2 E(x_{ij}^2)\right) \\ &= \frac{E(\delta_{ij}^2)}{E^2(x_{ij})} = \frac{\lambda_i^2}{E(\delta_{ij})} = \lambda_i^2\left[1 + \frac{e^{-\lambda_i T_s}(1 - e^{\lambda_i T_0})}{T_0\lambda_i}\right]^{-1}.\end{aligned}$$

Example 2.1 Arteriosclerotic Vascular Disease Trial

Cholesterol is the main lipid associated with arteriosclerotic vascular disease. The purpose of cholesterol testing is to identify patients at risk for arteriosclerotic heart disease. The liver metabolizes cholesterol to its free form and transports it to the bloodstream via lipoproteins. Nearly 75% of the cholesterol is bound to low-density lipoproteins (LDLs) – "bad cholesterol" and 25% is bound to high-density lipoproteins (HDLs) – "good

cholesterol." Therefore, cholesterol is the main component of LDLs and only a minimal component of HDLs and very low density lipoproteins. LDL is the substance most directly associated with increased risk of coronary heart disease (CHD).

Table 2.1: Sample Sizes for Different Types of Endpoints

Endpoint	Sample-Size	Variance
One mean	$n = \frac{(z_{1-a}+z_{1-\beta})^2\sigma^2}{\varepsilon^2}$;	
Two means	$n_1 = \frac{(z_{1-a}+z_{1-\beta})^2\sigma^2}{(1+1/r)^{-1}\varepsilon^2}$;	
One proportion	$n = \frac{(z_{1-a}+z_{1-\beta})^2\sigma^2}{\varepsilon^2}$;	$\sigma^2 = p(1-p)$
Two proportions	$n_1 = \frac{(z_{1-a}+z_{1-\beta})^2\sigma^2}{(1+1/r)^{-1}\varepsilon^2}$;	$\sigma^2 = \bar{p}(1-\bar{p})$; $\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1+n_2}$.
One survival curve	$n = \frac{(z_{1-a}+z_{1-\beta})^2\sigma^2}{\varepsilon^2}$;	$\sigma^2 = \lambda_0^2 \left(1 - \frac{e^{\lambda_0 T_0}-1}{T_0\lambda_0 e^{\lambda_0 T_s}}\right)^{-1}$
Two survival curves	$n_1 = \frac{(z_{1-a}+z_{1-\beta})^2\sigma^2}{(1+1/r)^{-1}\varepsilon^2}$;	$\sigma^2 = \frac{r\sigma_1^2+\sigma_2^2}{1+r}$, $\sigma_i^2 = \lambda_i^2 \left(1 - \frac{e^{\lambda_i T_0}-1}{T_0\lambda_i e^{\lambda_i T_s}}\right)^{-1}$

Note: $r = \frac{n_2}{n_1}$. λ_0= expected hazard rate, T_0= uniform patient accrual time and T_s= trial duration. Logrank-test is used for comparison of the two survival curves.

Suppose we are interested in a trial for evaluating the effect of a test drug on cholesterol in patients with CHD. A two-group parallel design is chosen for the trial with LDL as the primary endpoint. The treatment difference in LDL is estimated to be 5% with a standard deviation of 0.3. For power = 90% and one-sided $\alpha = 0.025$, the total sample can be calculated using (2.2):

$$n = \frac{4(1.96 + 1.28)^2\left(0.3^2\right)}{0.05^2} = 1212.$$

For a non-inferiority test, with a margin $\delta = -0.01$ (the determination of δ is a complicated issue and will not be discussed here.), the total sample-

size is given by

$$n = \frac{4(1.96 + 1.28)^2 \left(0.3^2\right)}{(0.05 + 0.01)^2} = 1050.$$

We can see that the required sample-size is smaller for the non-inferiority test than for a superiority test.

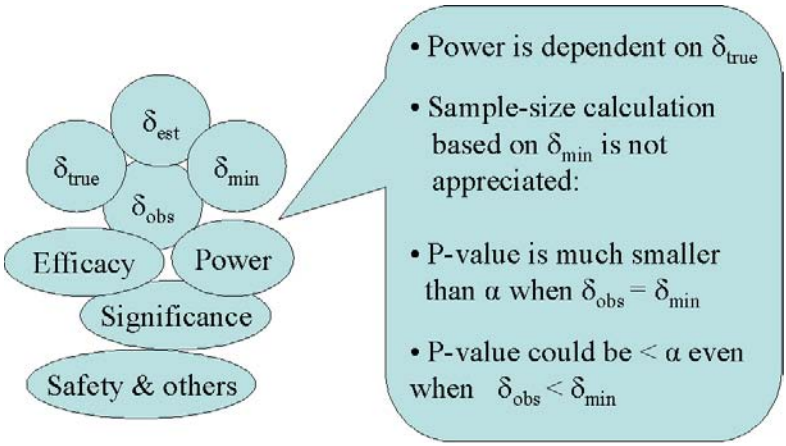


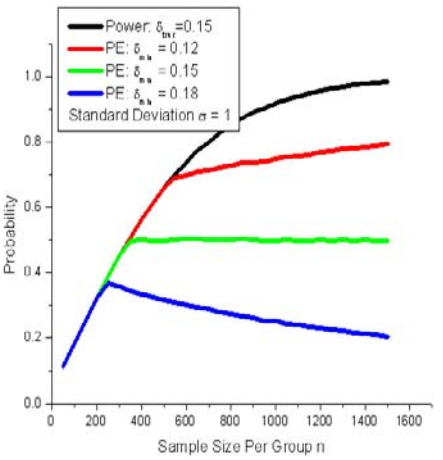
Figure 2.3: Sample-Size Calculation Based on δ

2.2.2 *Powering Trials Appropriately*

During the design, ε (δ_{true}) and σ are unknowns, but they can be estimated. Therefore, the power is just an estimation of the probability of achieving statistical significance and its precision is dependent on the precision of the initial estimate of ε and σ (Figure 2.3). When lacking information, a minimum clinically or commercially meaningful treatment difference $\delta_{\min}$ is often used. However, this strategy is not as good as it appears to be for the following reasons: (1) Power is not probability of success. The common phrase "90% power to detect a difference of $\delta_{\min}$" does not mean that there is a 90% probability of proving statistically that the treatment effect is larger than $\delta_{\min}$. What it really means is that if the true treatment difference is $\delta_{\min}$, then there is a 90% probability of proving a treatment difference > 0 (*zero*) at α level (Figure 2.4). (2) If the trial is designed based on $\delta_{\min}$, then as long as the observed treatment difference $\hat{\delta} > 0.61$ $\delta_{\min}$, there is a statistical significance (Figure 2.5). The trial is overpowered if the statistical significance is achieved even when there is no clinically or commercially meaningful magnitude of treatment effect. (3) If the true treatment difference is equal to $\delta_{\min}$, then there is 50% chance that the

observed treatment difference $\hat{\delta} > \delta_{\min}$ regardless of the sample-size (Figure 2.5). (4) $\delta_{\min}$ is difficult to know. Using the following formulation for the real superior design is too conservative:

$$n_2 = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2 (1 + 1/r)}{(\varepsilon - \delta_{\min})^2}. \tag{2.4}$$



- Increasing sample size does not make a drug better, but it improves the precision of the estimate.
- Statistical significance only proves treatment effect $\delta>0$
- Power = Pr ($p < \alpha$)
- Pe = Pr ($p < \alpha$ & $\delta_{obs} > \delta_{min}$) > Pr (Lower-CI-Limit > δ_{min})

δ_{min} = the minimal treatment difference with both clinical and commercial values.

Figure 2.4: Power and Probability of Efficacy (Pe)

The selections of the type-I error rate α and the type-II error rate β should be based on study objectives that may vary from phase to phase in clinical trials. It depends on efficacy, safety, and other aspects of the trial. From a safety perspective, the number of patients should be gradually increased from early phases to later phases due to the potential toxicity of the test drug. From an efficacy point-of-view, for early phases, there is more concern about missing good drug candidates and less concern about the false positive rate. In this case, a larger α is recommended. For later phases, a smaller α should be considered to meet regulatory requirements. In practice, it is suggested that the benefit-risk ratio should be taken into consideration when performing sample-size calculations. In such a case, Bayesian decision theory is a useful tool.

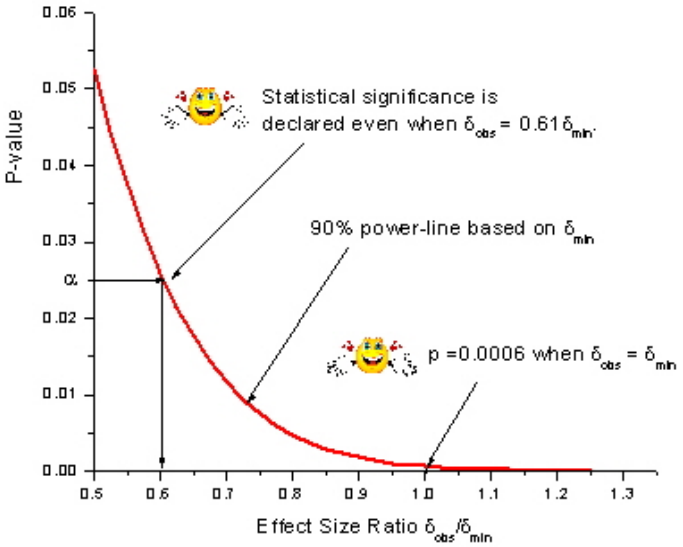


Figure 2.5: P-value Versus Observed Effect Size

2.3 Two-Group Equivalence Trial

2.3.1 *Equivalence Test*

The equivalence test for the two parallel groups can be stated as

$$H_0 : |\mu_T - \mu_R| \geq \delta \quad \text{versus} \quad H_a : |\mu_T - \mu_R| < \delta, \tag{2.5}$$

where the subscripts T and R refer to the test and reference groups, respectively. If the null hypothesis is rejected, then we conclude that the test drug and the reference drug are equivalent.

For a large sample-size, the null hypothesis is rejected if

$$T = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -z_{1-\alpha} \text{ and } T = \frac{\bar{x}_1 - \bar{x}_2 + \delta}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > z_{1-\alpha}. \tag{2.6}$$

The approximate sample-size is given by (Chow, Shao and Wang, 2003)

$$n_2 = \frac{\left(z_{1-\alpha} + z_{1-\beta/2}\right)^2 \sigma^2 \left(1 + 1/r\right)}{\left(|\varepsilon| - \delta\right)^2}, \tag{2.7}$$

where $r = n_1/n_2$.

Example 2.2 Equivalence LDL Trial

For the LDL trial in Example 2.1, assume the treatment difference $\varepsilon = 0.01$ and an equivalence margin of $\delta = 0.05$, the sample-size per group for a balanced design ($r = 1$) can be calculated using (2.7) with 90% power at $\alpha = 0.05$ level:

$$n_2 = \frac{(1.6446 + 1.6446)^2 \left(0.3^2\right) (1 + 1/1)}{(0.01 - 0.05)^2} = 1217.$$

Note that (2.7) is just an approximation even with a large sample-size, but an accurate calculation can be done using simulation. For a normal endpoint, the SAS Macro 2.1 can be used for power and sample-size calculations for equivalence studies. Note that the confidence interval method and the two one-sided tests method are equivalent. The SAS variables are defined as follows: **nSims** = number of simulation runs; **nPerGrp** = sample-size per group; **ux** = mean in group x; **uy** = mean in group y; **delta** = the equivalence margin; **sigmax** and **sigmay** = standard deviation for groups x and y, respectively; **alpha** = type-I error rate control; **xMean** and **yMean** = the simulated means in group x and y, respectively; **powerCI** = power based on the confidence interval method; and **powerTest** = power based on the two one-sided tests method. powerCI should be the same as powerTest.

>>SAS Macro 2.1: Equivalence Trial with Normal Endpoint>>

```
%Macro EquivCI(nSims=1000, nPerGrp=200, ux=0, uy=1, delta=1.2,
        sigmax=1, sigmay=1.2, alpha=0.05);
Data TwoGVars;
Keep xMean yMean powerCI powerTest;
powerCI=0; powerTest=0;
Do iSim=1 To &nSims;
 xMean=0; yMean=0; s2x=0; s2y=0;
 Do iObs=1 To &nPerGrp;
      xNOR=Rannor(7362); xMean=xMean+xNor; s2x=s2x+xNor**2;
      yNOR=Rannor(2637); yMean=yMean+yNor; s2y=s2y+yNor**2;
```

```
End;
xMean=xMean*&sigmax/&nPerGrp+&ux;
yMean=yMean*&sigmay/&nPerGrp+&uy;
sp=((s2x*&sigmax**2+s2y*&sigmay**2)/(2*&nPerGrp-2))**0.5;
se=sp/(&nPerGrp/2)**0.5;
* CI method;
ICW=Probit(1-&alpha)*se;
If Abs(yMean-xMean)+ICW < &delta Then
        powerCI=powerCI+1/&nSims;

*Two one-sided test method;
T1=(xMean-yMean-&delta)/se;
T2=(xMean-yMean+&delta)/se;
If T1<-Probit(1-&alpha) & T2>Probit(1-&alpha) Then
        powerTest=powerTest+1/&nSims;
End;
Output;
Run;
Proc Print Data=TwoGVars(obs=1); Run;
%Mend EquivCI;
```
‹‹SAS‹‹

The following SAS statements are examples of simulations under the null and alternative hypotheses.

››SAS››
```
Title "Equivalence test with Normal response: alpha under Ho";
%EquivCI(nSims=10000, nPerGrp=1000, ux=0.2, uy=0, delta=0.2,
sigmax=1, sigmay=1, alpha=0.05);

Title "Equivalence test with Normal response: Power under Ha";
%EquivCI(nSims=10000, nPerGrp=198, ux=0, uy=1, delta=1.2,
sigmax=0.8, sigmay=0.8, alpha=0.05);
```
‹‹SAS‹‹

For a binary endpoint, the power and sample-size for an equivalence test can be simulated using the SAS Macro 2.2. Note that the confidence interval method and the two one-sided tests method are equivalent. There definition of the SAS variables are defined as follows: **nSims** = number of simulation runs; **nPerGrp** = sample-size per group; **px** = response rate in group x; **py** = response rate in group y; **delta** = the equivalence margin; **sigmax** and **sigmay** = standard deviation for groups x and y, respectively;

alpha = type-I error rate control; **xMean** and **yMean** = the simulated means in group x and y, respectively; **powerCI** = power based on the confidence interval method; and **powerTest** = power based on the two one-sided tests method.

>>SAS Macro 2.2: Equivalence Trial with Binary Endpoint>>

```
%Macro TwoSamZTest(nSims=100000, nPerGrp=100,
                px=0.3, py=0.4, delta=0.3, alpha=0.05);
Data TwoGVars;
KEEP powerCI powerTest;
powerCI=0; powerTest=0;
Do iSim=1 To &nSims;
 PxObs=Ranbin(733,&nPerGrp,&px)/&nPerGrp;
 PyObs=Ranbin(236,&nPerGrp,&py)/&nPerGrp;
 se=((PxObs*(1-PxObs)+PyObs*(1-PyObs))/&nPerGrp)**0.5;
 *CI method;
 ICW=Probit(1-&alpha)*se;
 IF Abs(PxObs-PyObs)+ICW < &delta Then
                powerCI=powerCI+1/&nSims;
 *Two one-sided test method;
 T1=(PyObs-PxObs-&delta)/se;
 T2=(PyObs-PxObs+&delta)/se;
 IF T1<-Probit(1-&alpha) & T2>Probit(1-&alpha) Then
                powerTest=powerTest+1/&nSims;
End;
Output;
Run;
Proc Print; Run;
%Mend TwoSamZTest;
```

<<SAS<<

>>SAS>>

```
Title "Equivalence test with binary response: Alpha under Ho";
%TwoSamZTest(nPerGrp=100, px=0.1, py=0.2, delta=0.1, alpha=0.05);


Title "Equivalence test with binary response: Power under Ha";
%TwoSamZTest(nPerGrp=100, px=0.3, py=0.3, delta=0.2, alpha=0.05);
```

<<SAS<<

2.3.2 *Average Bioequivalence*

Pharmacokinetics (PK) is the study of the body's absorption, distribution, metabolism, and elimination of a drug. An important outcome of a PK study is the bioavailability of the drug. The bioavailability of a drug is defined as the rate and extent to which the active drug ingredient or therapeutic moiety is absorbed and becomes available at the site of drug action. As bioavailability cannot be easily measured directly, the concentration of drug that reaches the circulating bloodstream is taken as a surrogate. Therefore, bioavailability can be viewed as the concentration of drug that is in the blood. Two drugs are bioequivalent if they have the same bioavailability. There are a number of instances in which trials are conducted to show that two drugs are bioequivalent (Jones and Kenward, 2003): (1) when different formulations of the same drug are to be marketed, for instance in solid-tablet or liquid-capsule form; (2) when a generic version of an innovator drug is to be marketed, (3) when production of a drug is scaled up, and the new production process needs to be shown to produce drugs of equivalent strength and effectiveness as the original process.

At the present time, average bioequivalence (ABE) serves as the current international standard for bioequivalence (BE) testing using a 2×2 crossover design. The PK parameters used for assessing ABE are area under the curve (AUC) and peak concentration (Cmax). The recommended statistical method is the two one-sided tests procedure to determine if the average values for the PK measures determined after administration of the T (test) and R (reference) products were comparable. This approach is termed average bioequivalence (ABE). It is equivalent to the so-called confidence interval method, which involves the calculation of a 90% confidence interval for the ratio of the averages (population geometric means) of the measures for the T and R products. To establish BE, the calculated confidence interval should fall within a BE limit, usually $80\% - 125\%$ for the ratio of the product averages. The 1992 guidance has also provided specific recommendations for logarithmic transformation of PK data, methods to evaluate sequence effects, and methods to evaluate outlier data.

In practice, people also use parallel designs and the 90% confidence interval for nontransformed data. To establish BE, the calculated confidence interval should fall within a BE limit, usually $80\% - 120\%$ for the difference of the product averages.

The hypothesis for ABE in a 2×2 crossover design with log-transformed data can be written as

$$H_{01} : \mu_T - \mu_R \leq -\ln 1.25,$$
$$H_{02} : \mu_T - \mu_R \geq \ln 1.25.$$

The asymptotic power is given by (Chow, Shao, and Wang, 2003)

$$n = \frac{\left(z_{1-\alpha} + z_{1-\beta/2}\right)^2 \sigma_{1,1}^2}{2\left(\ln 1.25 - |\varepsilon|\right)^2},$$

where the variance for the intra-subject comparison is estimated using

$$\hat{\sigma}_{1,1}^2 = \frac{1}{n_1 + n_2 - 2} \sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(y_{i1j} - y_{i2j} - \bar{y}_{1j} + \bar{y}_{2j}\right)^2,$$

y_{ikj} is the log-transformed PK measure from the i^{th} subject in the j^{th} sequence at the k^{th} dosing period, and $\bar{y}_{kj}$ is the sample mean of the observations in the j^{th} sequence at the k^{th} period.

Example 2.3 Average Bioequivalence Trial

Suppose we are interested in establishing ABE between an inhaled formulation and a subcutaneously injected formulation of a test drug. The PK parameter chosen for this bioequivalence test is a log-transformation of the 24-hour AUC (i.e., the raw data is log-normal). Assume that the difference between the two formulations in log(AUC) is $\varepsilon = 0.04$ and the standard deviation for the intra-subject comparison is $\sigma_{1,1}^2 = 0.55$ with $\alpha = 0.05$ and $\beta = 0.2$, the sample-size per sequence is given by

$$n = \frac{(1.96 + 0.84)^2 (0.55)^2}{2(0.223 - 0.04)^2} = 36.$$

For a small sample, the bioequivalence test can be obtained using the following SAS macro Power2By2ABE. The purpose of this macro is to calculate sample-size for an average BE trial featuring a 2×2 crossover design with a normal endpoint. The power formulation was derived by Jones and Kenward (2003, p.336). The SAS variables are defined as follows: **sWithin** = Within-subject standard deviation on log-scale; **uRatio** = ratio of two treatment means; **n** = total sample-size; and **power** = power of the test.

››SAS Macro 2.3: Crossover Bioequivalence Trial››

```
%Macro Power2By2ABE(totalN=24, sWithin=0.355, uRatio=1);
```

```
Data ABE; Keep sWithin uRatio n power;
n=&totalN; sWithin=&sWithin; uRatio=&uRatio;
* Err df for AB/BA crossover design;
n2=n-2;
t1=tinv(1-0.05,n-2); t2=-t1;
nc1=Sqrt(n)*log(uRatio/0.8)/Sqrt(2)/sWithin;
nc2=Sqrt(n)*log(uRatio/1.25)/Sqrt(2)/sWithin;
df=Sqrt(n-2)*(nc1-nc2)/(2*t1);
Power=Probt(t2,df,nc2)-Probt(t1,df,nc1);
Run;
Proc Print; Run;
%Mend Power2By2ABE;
```

‹‹SAS‹‹

An example of how to use the macro is presend in the following:

››SAS››

```
%Power2By2ABE(totalN=58, sWithin=0.355, uRatio=1)
```

‹‹SAS‹‹

2.3.3 *Population and Individual Bioequivalence*

An FDA 2001 guidance describes two new approaches, termed population bioequivalence and individual bioequivalence (PBE, IBE). PBE is concerned with assessing if a patient who has not yet been treated with R or T can be prescribed either formulation. IBE is a criterion for deciding if a patient who is currently being treated with R can be switched to T. The ABE method does not assess a subject-by-formulation interaction variance, that is, the variation in the average T and R difference among individuals. In contrast, PBE and IBE approaches include comparisons of both averages and variances of the measure. The PBE approach assesses total variability of the measure in the population. The IBE approach assesses within-subject variability for the T and R products, as well as the subject-by-formulation interaction. For PBEs and IBEs, the 95% confidence intervals are recommended with the same BE limits as those for ABE.

Statistical analyses of PBE and IBE data typically require a higher-order crossover design such as [RTR,TRT] or [RTRT,TRTR]. The statistical model is often a mixed-effects model. PBE and IBE approaches, but not the ABE approach, allow two types of scaling: reference scaling and constant scaling. Reference scaling means that the criterion used is scaled to the variability of the R product, which effectively widens the BE limit for more

variable reference products. Although generally sufficient, use of reference scaling alone could unnecessarily narrow the BE limit for drugs and/or drug products that have low variability but a wide therapeutic range. This guidance, therefore, recommends mixed scaling for the PBE and IBE approaches. With mixed scaling, the reference-scaled form of the criterion should be used if the reference product is highly variable; otherwise, the constant-scaled form should be used.

The hypothesis test for IBE is given by

$$H_0 : \begin{cases} (\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - 3.49\sigma_{WR}^2 \geq 0 & \text{if } \sigma_{WR}^2 > 0.04, \\ (\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2 - 0.996 \geq 0 & \text{if } \sigma_{WR}^2 \leq 0.04, \end{cases}$$

where σ_{WT}^2 and σ_{WR}^2 are the within-subject variances for T and R, respectively, σ_{BT}^2 and σ_{BR}^2 are the between-subject variances for T and R, and $\sigma_D^2 = \sigma_{BT}^2 + \sigma_{BR}^2 - 2\rho\sigma_{BT}^2\sigma_{BR}^2$ is the subject-by-formulation interaction, where ρ is the between-subject correlation of T and R. The mean of T and R are denoted by μ_T and μ_R, respectively.

The hypothesis test for PBE is given by

$$H_0 : \begin{cases} (\mu_T - \mu_R)^2 + \sigma_T^2 - 3.49\sigma_R^2 \geq 0 & \text{if } \sigma_R^2 > 0.04, \\ (\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2 - 0.996 \geq 0 & \text{if } \sigma_R^2 \leq 0.04, \end{cases}$$

where $\sigma_T^2 = \sigma_{WT}^2 + \sigma_{BT}^2$ and $\sigma_R^2 = \sigma_{WR}^2 + \sigma_{BR}^2$.

Further details can be found in (Jones and Kenward, 2003, and Chow and Liu, 2003). SAS programs for IBE and PBE are available from Jones and Kenward (2003).

2.4 Dose-Response Trials

Dose-response trials are also called dose-finding trials. Four questions are often of interest in a dose-response trial (Ruberg, 1995): (1) Is there any evidence of drug effect? (2) What doses exhibit a response different from the control response? (3) What is the nature of the dose-response? and (4) What is the optimal dose? A phase-II dose-response trial is typically a multiple-arm parallel design with a control group. There are a variety of approaches to statistical analysis for a dose response study; for examples, see Chuang and Agresti (1997) and Stewart and Ruberg (2000). A commonly used and conservative approach is to compare each active dose to the control using Dunnett's test or a stepwise test. As pointed out by

Steart and Ruberg (2000), the contrast will detect certain expected dose-response features without forcing those expected features into the analysis model. Commonly used contrast procedures include Dunnett's test (Dunnett, 1955), the regression test of Tukey et al. (Tukey and Ciminera, 1885), Ruberg's basin contrast (Ruberg, 1989), Williams's test (Williams, 1971, 1972), and the Cochran-Armitage test (Cochran, 1954; Amitage, 1955). For multiple contrast tests, there are usually multiplicity adjustment requirements (Hsu and Berger, 1999). The sample-size formulation is available for multi-arm dose-response trials for binary endpoints based on contrast tests (Nam, 1987). For ordered categorical data, Whitehead (1993) derived a formulation for sample-size calculation based on a proportional-odds model.

The objective of this section is to provide a unified formulation and a user-friendly SAS macro for calculating the power and sample-size for multiple-arm superiority and noninferiority trials with continuous, binary, or survival endpoints (Chang, 2006; Chang, and Chow, 2006a).

2.4.1 *Unified Formulation for Sample-Size*

In multiple-arm trials, a general one-sided hypothesis testing problem can be stated as a contrast test:

$$H_o : L(\mathbf{u}) \leq 0; \text{ vs. } H_a : L(\mathbf{u}) = \varepsilon > 0, \tag{2.8}$$

where the operator or function $L(\cdot)$ is often linear, $\mathbf{u} = \{u_i\}$, u_i can be the mean, proportion, or hazard rate for the i^{th} group depending on the study endpoint, and ε is a constant.

A test statistic can be defined as

$$T = \frac{L(\hat{\mathbf{u}})}{\sqrt{var_{\varepsilon=0}(L(\hat{\mathbf{u}}))}}, \tag{2.9}$$

where $\hat{\mathbf{u}}$ is an unbiased estimator of $\mathbf{u}$.

A linear operator of $L(\cdot)$ is particularly interesting and will be used in the rest of the chapter:

$$L(\mathbf{u}) = \sum_{i=1}^{k} c_i u_i - \delta, \tag{2.10}$$

where the contrast coefficient c_i satisfies the equation $\sum_{i=1}^{k} c_i = 0$ ($c_1 = 1$ for a single-arm trial). Without losing generality, assume that $c_i u_i > 0$ indicates efficacy; then, for a superiority design, $\delta \geq 0$, and for a noninferiority

design, $\delta < 0$. Note that if $\delta = 0$ and H_o defined by (2.8) is rejected for some $\{c_i\}$ satisfying $\sum_{i=1}^{k} c_i = 0$, then there is a difference among u_i $(i = 1, ..., k)$.

Let $\hat{\mathbf{u}}$ be the mean for a continuous endpoint, proportion for a binary endpoint, and maximum likelihood estimator (MLE) of the hazard rate for a survival endpoint; then, the asymptotic distributions of the test statistic can be obtained from the central limit theorem:

Under the null hypothesis, the test statistic is given by

$$T = \frac{L_{\varepsilon=0}(\hat{\mathbf{u}})}{v_o} \sim N(0, 1) \tag{2.11}$$

and under the specific alternative hypothesis associated with ε, the test statistic is given by

$$T = \frac{L(\hat{\mathbf{u}})}{v_o} \sim N(\frac{\varepsilon}{v_o}, \frac{v_a^2}{v_o^2}), \tag{2.12}$$

where

$$\varepsilon = E(L(\hat{\mathbf{u}}), \tag{2.13}$$

$$\begin{cases} v_o^2 = var_{\varepsilon=0}(L(\hat{\mathbf{u}})) \\ v_a^2 = var(L(\hat{\mathbf{u}})) \end{cases} . \tag{2.14}$$

Because of (2.10), (2.14) can be written as

$$\begin{cases} v_o^2 = \sum_{i=1}^{k} c_i^2 var_{\varepsilon=0}(\hat{u}_i) = \sigma_o^2 \sum_{i=1}^{k} \frac{c_i^2}{n_i} = \frac{\theta_o^2}{n} \\ v_a^2 = \sum_{i=1}^{k} c_i^2 var(\hat{u}_i) = \sum_{i=1}^{k} \frac{c_i^2 \sigma_i^2}{n_i} = \frac{\theta_a^2}{n} \end{cases} , \tag{2.15}$$

where

$$\begin{cases} \theta_o^2 = \sigma_o^2 \sum_{i=1}^{k} \frac{c_i^2}{f_i} \\ \theta_a^2 = \sum_{i=1}^{k} \frac{c_i^2 \sigma_i^2}{f_i} \end{cases} , \tag{2.16}$$

where n_i is the sample-size for the i^{th} arm, $f_i = \frac{n_i}{n}$ is the size fraction, $n = \sum_{i=0}^{k} n_i$, σ_o^2 is the variance of the response under H_o, and σ_i^2 is the variance under H_a for the i^{th} arm.

From (2.12) and (2.15), it is immediately obtained that under the specific alternative hypothesis, the test statistic T is normally distributed with a mean of $\frac{\sqrt{n}\varepsilon}{\theta_o}$ and a variance of $\frac{\theta_a^2}{\theta_o^2}$. Therefore, similar to (2.1), the power considering heterogeneity of variances can be obtained:

$$\text{power} = \Phi\left(\frac{\varepsilon\sqrt{n} - \theta_o z_{1-\alpha}}{\theta_a}\right). \tag{2.17}$$

Similar to (2.2), the sample-size with the heterogeneous variances is given by

$$n = \frac{(z_{1-a}\theta_o + z_{1-\beta}\theta_a)^2}{\varepsilon^2}. \tag{2.18}$$

Note that ε defined by (2.13) is the treatment difference Δ - the non-inferior/superiority margin δ. When $\delta = 0$, ε is simply the treatment difference.

Equations (2.16) through (2.18) are applicable to any k-arm design ($k \geq 1$). The asymptotic variance σ_i^2 can be estimated by

$$\hat{\sigma}_i^2 = \hat{p}_i(1 - \hat{p}_i) \tag{2.19}$$

for a binary endpoint with an estimated response rate of $\hat{p}_i$, and

$$\hat{\sigma}_i^2 = \hat{\lambda}_i^2 \left[1 + \frac{e^{-\hat{\lambda}_i T_s}(1 - e^{\hat{\lambda}_i T_0})}{T_0 \hat{\lambda}_i}\right]^{-1} \tag{2.20}$$

for an exponentially distributed survival endpoint with an estimated hazard rate of $\hat{\lambda}_i$. These two variances can be used to calculate θ_o^2 and θ_a^2 in (2.16) when the sample-size is large. It can be seen that (2.17) and (2.18) have included the common one-arm and two-arm superiority and noninferiority designs as special cases: for a one-arm design, $c_1 = 1$, and for a two-arm design, $c_1 = -1$ and $c_2 = 1$.

2.4.2 *Application Examples*

Three examples (all modified from the actual trials) will be used to demonstrate the utility of the proposed method for clinical trial designs. The first example is a multiple-arm trial with a continuous endpoint. In the second example, both superiority and noninferiority designs are considered for a multiple-arm trial with a binary endpoint. The third example is an application of the proposed method for designing a multiple-arm trial with a survival endpoint, where different sets of contrasts and balanced, as well as unbalanced, designs are compared. For convenience, the SAS macro for sample-size calculation is provided.

Example 2.4 Dose-Response Trial with Continuous Endpoint

In a phase II asthma study, a design with 4 dose groups (0 mg, 20 mg, 40 mg, and 60 mg) of the test drug is proposed. The primary efficacy endpoint

is the percent change from baseline in forced expiratory volume in the first second (FEV1). From previous studies, it has been estimated that there will be 5%, 12%, 13%, and 14% improvements over baseline for the control, 20 mg, 40 mg, and 60 mg groups, respectively, and a homogeneous standard deviation of $\sigma = 22\%$ for the FEV1 change from baseline. To be consistent with the response shape, let the contrast $c_i = 100(\mu_i - \frac{1}{4}\sum_{i=1}^{4}\mu_i)$, i.e., $c_1 = -6$, $c_2 = 1$, $c_3 = 2$, $c_4 = 3$, where μ_i is the estimated FEV1 improvement in the i^{th} group. It can be seen that any set of contrasts with multiples of the above $\{c_i\}$ will lead to the same sample-size. Thus it can be obtained that $\varepsilon = \sum_{i=1}^{4} c_i\mu_i = 50\%$. Using a balanced design ($f_i = 1/4$) with a one-sided $\alpha = 0.05$, the sample-size required to detect a true difference of $\varepsilon = 0.5$ with 80% power is given by

$$\begin{aligned} n &= \left[\frac{(z_{1-\alpha} + z_{1-\beta})\sigma}{\varepsilon}\right]^2 \sum_{i=1}^{4} \frac{c_i^2}{f_i} \\ &= \left[\frac{(1.645 + 0.842)(0.22)}{0.50}\right]^2 4((-6)^2 + 1^2 + 2^2 + 3^2) \\ &= 240. \end{aligned}$$

Thus, a total sample-size of 240 is required for the trial.

Example 2.5 Dose-Response Trial with Binary Endpoint

A trial is to be designed for patients with acute ischemic stroke of recent onset. The composite endpoint (death and miocardial infarction [MI]) is the primary endpoint. There are four dose levels planned with event rates of 14%, 13%, 12%, and 11%, respectively. The first group is the active control group (14% event rate). It is interested in both superiority and noninferiority tests comparing the test drug to the active control. Notice that there is no need for multiplicity adjustment for the two tests because of the closed-set test procedure. The comparisons are made between the active control and the test groups; therefore, the contrast for the active control should have a different sign than the contrasts for the test groups. Let $c_1 = -6$, $c_2 = 1$, $c_3 = 2$, and $c_4 = 3$. It is assumed that the noninferiority margin for the event rate is $\delta = 0.5\%$, and the event rate is $p_o = 0.14$ under the null hypothesis. Because it is a noninferiority design and the noninferiority margin is usually defined based on a two-arm design, to make this noninferiority margin usable in the multiple-arm design, the contrasts need to be rescaled to match the two-arm design, i.e., set the contrast for the control group $c_1 = -1$. The final contrasts used in the trial are given by $\left\{c_1 = -1,\ c_2 = \frac{1}{6},\ c_3 = \frac{1}{3},\ c_4 = \frac{1}{2}\right\}$. Based on this information, it can

be obtained that $\varepsilon = \sum_{i=1}^{k} c_i p_i - \delta = -0.02333 - 0.005 = -0.02833$, where p_i is the estimated event rate in the i^{th} group. Using a balanced design ($f_i = 1/4$), the two key parameters, θ_o^2 and θ_a^2, can be calculated as follows:

$$\begin{cases} \theta_o^2 = p_o(1-p_o)\sum_{i=1}^{k} \frac{c_i^2}{f_i} = 0.6689 \\ \theta_a^2 = \sum_{i=1}^{k} \frac{c_i^2\ p_i(1-p_i)}{f_i} = 0.639 \end{cases}.$$

Using a one-sided $\alpha = 0.025$ and a power of 90%, the sample-size required for the noninferiority test is given by

$$\begin{aligned} n &= \left[\frac{(z_{1-\alpha}\theta_o + z_{1-\beta}\theta_a)}{\varepsilon}\right]^2 \\ &= \left[\frac{(1.96\sqrt{0.6689} + 1.2815\sqrt{0.639})}{-0.02833}\right]^2 \\ &= 8600. \end{aligned}$$

Thus a total sample-size of 8600 patients is required for the noninferiority test. With 8600 patients, the power for the superiority test ($\delta = 0, \varepsilon = 0.0233$) is 76.5%, which is calculated as follows:

$$\begin{aligned} \text{power} &= \Phi_o\left(\frac{\varepsilon\sqrt{n} - \theta_o z_{1-\alpha}}{\theta_a}\right) \\ &= \Phi_o\left(\frac{0.0233\sqrt{8600} - 1.96\sqrt{0.6689}}{\sqrt{0.639}}\right) \\ &= \Phi_o(0.6977) = 76\%. \end{aligned}$$

Note that different contrasts can be explored to minimize the sample-size.

An interesting note is that the Cochran-Armitage linear trend test is a special case of the contrast test in which the contrast $c_i = d_i - \bar{d}$, where d_i is the i^{th} dose and $\bar{d}$ is the average dose.

Example 2.6 Dose-Response Trial with Survival Endpoint

Let λ_i be the population hazard rate for group i. The contrast test for multiple survival curves can be written as $H_o : \sum_{i=0}^{k} c_i\lambda_i \leq 0$. This null hypothesis is assumed in the following example.

In a four-arm (active control, lower dose of test drug, higher dose of test drug, and combined therapy), phase II oncology trial, the objective is to determine if there is treatment effect with time-to-progression as the primary endpoint. Patient enrollment duration is estimated to be $T_0 = 9$ months and the total trial duration is $T_s = 16$ months. The estimated

median times for the four groups are 14, 20, 22, and 24 months with the corresponding hazard rates of 0.0459, 0.0347, 0.0315, and 0.0289/month, respectively (under the exponential survival distribution, $\lambda T_{Median} = ln2$). The hazard rate under the null hypothesis is assumed to be 0.03525. A power of 80% and a one-sided α of 0.025 are proposed for the trial. The small α is used due to the consideration of potential accelerated approval using this trial. In order to achieve the most efficient design (i.e., minimum sample-size), sample sizes from different contrasts and various designs (balanced and unbalanced) are compared. The results are presented in Table 2.2, where the optimal design is the minimum variance design in which the number of patients assigned to each group is proportional to the variance of the group. It can be seen that the optimal design with sample-size fractions (0.343, 0.244, 0.217, and 0.197) is generally the most powerful and requires fewer patients regardless of the shape of the contrasts. The contrasts with a linear trend also work well for the optimal design. Although the optimal design with linear contrasts seems attractive with a total sample-size of 646 subjects, in practice, more patients being assigned to the control group presents an ethical concern, and it is desirable to obtain more information on the test groups. Therefore, a balanced design with contrasts (10.65, -0.55, -3.75, and -6.35) is recommended with a total sample-size of 742 subjects.

Table 2.2: Sample Sizes for Different Contrasts (Balanced Design)

Scenario	Contrast				Total n	
					Balance	Optimal
Average dose effect	-3	1	1	1	838	690
Linear response trend	-6	1	2	3	759	646
Median time trend	-6	0	2	4	742	664
Hazard-rate trend	10.65	-0.55	-3.75	-6.35	742	651

Note: Sample-size fractions for the optimal design = 0.343, 0.244, 0.217, and 0.197.

2.4.3 *Determination of Contrast Coefficients*

There are two criteria that need to be considered when selecting contrasts: (1) The selected contrasts must lead to a clinically meaningful hypothesis test, and (2) The selected contrasts should provide the most powerful test statistic after criterion 1.

To use a contrast test, the selection of contrasts should be practically meaningful. If one is interested in a treatment difference among any groups, then any contrasts can be applied. If one is only interested in the compar-

ison between dose-level 1 and other dose levels, then one should make the contrast for dose-level 1 have a different sign from that of the contrasts for other dose groups. Otherwise, efficacy may not be concluded even when the null hypothesis H_o is rejected, because the rejection of H_o could be due simply to the opposite effects (some positive and some negative) of different dose levels of the test drug.

To study how the different combinations of response shapes and contrasts may affect the sample-size and power, the following five different shapes (Table 2.3) are considered.

Table 2.3: Response and Contrast Shapes

Shape	u_1	u_2	u_3	u_4	c_1	c_2	c_3	c_4
Linear	0.1	0.3	0.5	0.7	-3.00	-1.00	1.00	3.00
Step	0.1	0.4	0.4	0.7	-3.00	0.00	0.00	3.00
Umbrella	0.1	0.4	0.7	0.5	-3.25	-0.25	2.75	0.75
Convex	0.1	0.1	0.1	0.6	-1.25	-1.25	-1.25	3.75
Concave	0.1	0.6	0.6	0.6	-3.75	1.25	1.25	1.25

Note: $c_i = b\left(u_i - \frac{1}{4}\Sigma_{i=1}^{4} u_i\right)$, b = any constant.

Sample sizes required under a balanced design for different combinations of responses and contrasts are presented in Table 2.4. It can be seen that under a balanced design, when response and contrasts have the same shape, a minimal sample-size is required. If an inappropriate contrast set is used, the sample-size could be 30 times larger than the optimal design.

Table 2.4: Sample-Size Per Group for Various Contrasts

	Contrast				
Response	Linear	Step	Umbrella	Convex	Concave
Linear	31	35	52	52	52
Step	39	35	81	52	52
Umbrella	55	74	33	825	44
Convex	55	50	825	33	297
Concave	55	50	44	297	33

Note: $\sigma = 1$, one-sided $\alpha = 0.05$

In fact, under a balanced design, homogenous variance under H_o and H_a and $\delta = 0$, the minimum sample-size or maximum power is achieved when the following equation is satisfied (assume $\bar{u} = \sum_{i=1}^{k} u_i = 0$):

$$\frac{\partial n}{\partial c_i} = 0. \tag{2.21}$$

Under the given conditions, (2.21) is equivalent to

$$\frac{\partial}{\partial c_i}\left(\frac{\Sigma_{i=1}^{M} c_i^2}{\left[\Sigma_{i=1}^{M} c_i u_i\right]^2}\right) = 0. \tag{2.22}$$

It is obvious that the solution to (2.22) is $c_i = u_i$ $(i = 1, ..., k)$. If $\bar{u} \neq 0$, we can make a linear transformation $u_i^* = u_i - \bar{u}$; hence $c_i = u_i^*$ or $c_i = u_i - \bar{u}$ for minimum sample-size.

2.4.4 *SAS Macro for Power and Sample-Size*

For convenience, the sample-size calculation formulation (2.18) has been implemented in SAS macro AsympN. This SAS macro can be used to calculate the sample-size for multiple-arm superiority and noninferiority trial designs with continuous, binary, or survival endpoints. The parameters are defined as follows: **endpoint** = "normal", "binary", or "survival" ; **alpha** = one-sided significance level; **nArms** = number of groups; **delta** (> 0) = superiority margin, and **delta** (<0) = in-inferiority margin; **tAcr** = patient enrollment duration; **tStd** = study duration; **u{i}** are treatment mean, proportions of response, or hazard rates for the i^{th} group; **s{i}** = standard deviations for a continuous endpoint; **c{i}** = the contrasts; and **f{i}** = sample-size fractions among treatment groups. Note that **tAcr** and **tStd** are for a survival endpoint only, and Σ**c{i}**=0. The standard deviation under the null hypothesis is assumed to be the average standard deviation over all groups.

>>SAS Macro 2.4: Sample-Size for Dose-Response Trial>>

```
%Macro AsympN(endpoint="normal", alpha=0.025, power=0.8,
                    nArms=5, delta=0, tStd=12, tAcr=4);
Data contrastN; Set dInput;
Keep Endpoint nArms alpha power TotalSampleSize;
Array u{&nArms}; Array s{&nArms}; Array f{&nArms};
Array c{&nArms}; endpoint=&endpoint; delta=&delta;
alpha=&alpha; power=&power; nArms=&nArms;
 epi = 0; s0 = 0;
 Do i =1 To nArms; epi = epi + c{i}*u{i}- &delta; End;
 If &endpoint = "normal" Then Do;
     Do i =1 To nArms; s0 = s0 + s{i}/nArms; End;
 End;
 If &endpoint = "binary" Then Do;
     Do i = 1 To nArms;
```

```
            s{i} = (u{i}*(1-u{i}))**0.5;
            s0=s0 + s{i}/nArms;
        End;
    End;
    If &endpoint = "survival" Then Do;
        Do i = 1 To nArms;
            s{i} = u{i}*(1+exp(-u{i}*&tStd)*(1-exp(u{i}*&tAcr))
                /(&tAcr*u{i}))**(-0.5);
            s0 = s0 + s{i}/nArms;
        End;
    End;
    sumscf0 = 0; sumscf = 0;
    Do i = 1 To nArms;  sumscf0 = sumscf0 + s0**2*c{i}*c{i}/f{i}; End;
    Do i = 1 To nArms;  sumscf = sumscf + s{i}**2*c{i}*c{i}/f{i}; End;
    n = ((PROBit(1-&alpha)*sumscf0**0.5
            + Probit(&power)*sumscf**0.5)/epi)**2;
    TotalSampleSize = round(n);
run;
proc print;
run;
%Mend AsympN;
```

‹‹SAS‹‹

The following example shows how to call this SAS macro for sample-size calculations with normal, binary, and survival endpoints.

››SAS››

```
Title " = s of How to Use the SAS Macros";
Data dInput;
Array u{4}(.46, .35, .32, .3);      ** Responses;
Array s{4}(2, 2, 2, 2);          ** Standard deviation for normal endpoint;
Array c{4}(-4, 1, 1, 2);                ** Contrasts;
Array f{4} (.25, .25 ,.25, .25);    ** Sample size fractions;
%AsympN(endpoint="normal", alpha=0.025, power=0.8, nArms=4);
%AsympN(endpoint="binary", alpha=0.025, power=0.8, nArms=4);
%AsympN(endpoint="survival", alpha=0.025, power=0.8, nArms=4,
            delta=0, tStd=2, tAcr=.5);
run;
```

‹‹SAS‹‹

2.5 Maximum Information Design

In clinical trials, the sample-size is determined by a clinically meaningful difference and information on the variability of the primary endpoint. Due to lack of knowledge of the new treatment, estimates of the variability for the primary endpoint are often not precise. As a result, the initially planned sample-size may turn out to be inappropriate and needs to be adjusted at interim analysis to ensure the power if the observed variability of the accumulated response on the primary endpoint is very different from that used at the planning stage. To maintain the integrity of the trial, it is suggested that sample-size re-estimation be performed without unblinding the treatment codes if the study is to be conducted in a double-blind fashion. Procedures have been proposed for adjusting the sample-size during the course of the trial without unblinding and altering the significance level (Gould, 1992; Gould and Shih, 1992). An alternative approach to dealing with the noise of the pooled variance is to use the maximum information design. The idea behind this approach is that recruitment continues until the prespecified information level ($I = N/(2\hat{\sigma}^2) = I_{\max}$) is reached. For a given sample-size, the information level is reduced when the observed variance increases. The total sample-size for the two-group parallel design can be written in this familiar form:

$$N = \frac{4}{\delta^2}(z_{1-\alpha/2} + z_{1-\beta})^2 \frac{\sigma_0^2}{\hat{\sigma}^2}, \tag{2.23}$$

where δ is treatment difference and σ_0 is the estimated standard deviation at the time of designing the trial and $\hat{\sigma}$ is the observed standard deviation.

If $\hat{\delta}$ is not related to $\hat{\sigma}$, then N is independent of $\hat{\delta}$. Adjusting the N based on $\hat{\sigma}$ will not inflate the overall α.

2.6 Summary and Discussion

In this chapter, we reviewed commonly used classic trial design methods. The methods derived from a contrast test can be used for power and sample-size calculations for k-arm trials ($k \geq 1$). They can be used for superiority or noninferiority designs with continuous, binary, or survival endpoints. The selection of contrasts is critical. The selected contrasts must lead to a clinically meaningful hypothesis test and should lead to a powerful test statistic. The examples above provided details about the use of these methods. Contrast testing can be used to detect treatment difference. The most (or least) responsive arm can be considered superior or noninferior to

the control and can be selected for studies in the next phase. As far as the response shape is concerned, model approaches or multiple-contrast tests can be used to establish the confidence intervals or predictive intervals of the response for each dose level under study. The optimal dose is more complicated because the safety aspect has to be considered.

We have also discussed equivalence studies and the maximum information design. The latter can be considered a quasi adaptive design because the sample-size changes automatically based on the nuisance parameter σ^2.

It is important to remember that the power very much relies on the assumption of the estimated effect size at the time of study design. It is even more critical to fully understand these three different concepts about effect size: true size, estimated size, and minimum meaningful effect size, and their impacts on trial design. Last but not least, trial design involves many different aspects of medical/scientific, statistical, commercial, regulatory, operational, data management functions. A statistician cannot view the achievement of a design with the greatest power or smallest sample-size as the ultimate goal. Further, a trial design cannot be viewed as an isolated task. Instead, drug development should be viewed as an integrated process in which a sequence of decisions are made. We will discuss more on this throughout the book.

Problems

2.1 A clinical design team is discussing the trial design for a phase-III asthma study. Based on the results of a phase-II trial, the percent increase from baseline in FEV is 6%, 11%, and 15% for placebo, 400 mg, and 800 mg dose groups, respectively. The common standard deviation is 18%. No safety concerns have been raised based on the phase-II data. The medical research and commercial groups in the company believe that the clinically and commercially meaningful minimum treatment difference is 7% between active group and placebo because a commercial product with 7% mean FVE1 improvement over placebo is available on the market with a good safety profile.

Design this phase-III trial (type of design, number of groups and dose levels, sample-size). Justify your design, determine the p-value if the observed is 6.9%, 7%, and 7.1%, and discuss the implications of these p-values to your design.

2.2 An Oncology Trial Design. Design the following trial and recommend a sample-size.

Consider a two-arm oncology trial comparing a test treatment to an active control with respect to the primary efficacy endpoint, time to disease progression (TTP). Based on data from previous studies, the median TTP is estimated to be 10 months (hazard rate = 0.0693) for the control group, and 13 months (hazard rate = 0.0533) for the test treatment group. Assume that there is a uniform enrollment with an accrual period of 10 months and that the total study duration is expected to be 24 months.

2.3 Some Commonly Used Formulas

(1) For PK/PD and Bioequivalence studies, log-transformation is often used.

(a) Prove the following:

$$\sigma_{\ln X} = \sqrt{\ln\left(1 + CV_X\right)} \text{ if } \ln X \text{ is Normal.}$$

(b) Prove the following relationship under a general condition with a small CV:

$$\ln X \simeq \ln \mu + \frac{X - \mu}{\mu} \sim N\left(\ln \mu,\ CV^2\right).$$

(2) For a survival analysis, the power is often based on the number of events instead of the number of patients. Therefore, we can use the exponential model to predict the time when a certain number of events is reached. This is very useful for operational planning. Prove the following relationships under the assumption of exponential distribution:

$$D = \begin{cases} R\left(T - \frac{1}{\lambda} + \frac{1}{\lambda}e^{-\lambda T}\right) & \text{if } T \le T_0, \\ R\left[T_0 - \frac{1}{\lambda}\left(e^{\lambda T_0} - 1\right)e^{-\lambda T}\right] & \text{if } T > T_0, \end{cases}$$

and

$$T = \begin{array}{ll} -\frac{1}{\lambda}\ln\left(\frac{\lambda D}{R} - \lambda T + 1\right) & \text{if } T \le T_0, \\ -\frac{1}{\lambda}\ln\left[\lambda\left(T_0 - \frac{D}{R}\right)\left(e^{\lambda T_0} - 1\right)^{-1}\right] & \text{if } T > T_0 \end{array},$$

where T_0 = enrollment duration, T = the time of interesting from randomization, D = number of deaths, R = uniform enrollment rate, λ = hazard rate.

Also prove the following under exponential distribution:

$$T_{median} = \frac{\ln 2}{\lambda} = T_{mean} \ln 2$$

and the two-sided $(1-\alpha)\%$ confidence interval for the hazard λ:

$$\left[\frac{\hat{\lambda}}{2D}\chi^2_{2D,\ 1-\alpha/2},\ \frac{\hat{\lambda}}{2D}\chi^2_{2D,\ \alpha/2}\right],$$

where T_{median} = median time, T_{mean} = mean time, $\hat{\lambda}$ = MLE of λ.

2.4 Power and Sample-Size Formulation for a Model-Based Approach to Dose-Response Trials

Test-based approaches are fine for detecting evidence against the null hypothesis in the direction of a positive trend. However, they do not provide much insight into the form of the relationship. A model-based perspective is better for this purpose. A good-fitting model describes the nature of the association, provides parameters for describing the strength of the relationship, provides predicted probabilities for the response categories at any dose, and helps us to determine the optimal dose. It also yields the hypothesis of no treatment effect if a frequentist approach is used for the modeling. However, the results from model-based approaches are heavily dependent on the accuracy of the model to the natural phenomenon.

Whitehead (Whitehead, 1993, and Chuang and Agresti, 1997) developed the sample-size formulation for an ordinal response based on the proportional odds model (logistic model) for two groups. The total sample-size for a one-sided test is given by

$$N = \frac{2(r+1)^2(z_{1-\alpha} + z_{1-\beta})^2}{r(\ln R)^2(1 - \sum \bar{p}_i^3)},$$

where r is sample-size ratio, R is odd ratio which can be obtained from logistic regression with or without covariates, and $\bar{p}_i$ is the anticipated marginal proportion in the response category i.

The power is given by

$$power = \Phi\left(\sqrt{\frac{N\, r\, (\ln R)^2(1 - \sum \bar{p}_i^3)}{2(r+1)^2}} - z_{1-\alpha}\right).$$

Generalize the formulations for sample-size and power for dose-response models other than logistic model.

2.5 Reproduction

Consider a trial with two parallel arms comparing the mean difference. Assume that the known variance $\sigma^2 = 1$ and the true treatment difference is δ. The estimated treatment difference is δ_0. The trial was design at level α with $(1-\beta)$ power and sample-size n. In other words, $P_{\delta=\delta_0}(p \leq \alpha) = \alpha$, $P_{\delta=\delta_0}(p \leq \alpha) = 1 - \beta$. If the observed treatment difference $\hat{\delta}$ at the end of the trial is less than, equal to, or larger than the true difference δ, what is in each case the probability (reproductivity) that the next trial show statistical significance with sample-size n. Can we use the reproductivity of 50% when $\delta = \hat{\delta}$ to argue that $\alpha = 0.025\%$ is too unconservative? Why?

2.6 Correlation between response difference and common variance

Equation (2.23) is valid when $\hat{\delta}$ is not related to $\hat{\sigma}$, Draw a function (or graphically via simulation) to reveal the relationship between $\hat{\delta}$ and $\hat{\sigma}$ for Normal, binary and survival endpoint with finite sample-size.

Chapter 3

Theory of Adaptive Design

3.1 Introduction

As indicated early in Chapter 1, an adaptive design is a design that allows adaptations or modifications to some aspects of a trial after its initiation without undermining the validity and integrity of the trial. The adaptations may include, but are not limited to, sample-size re-estimation, early stopping for efficacy or futility, response-adaptive randomization, and dropping inferior treatment groups (Figure 3.1). Adaptive designs usually require unblinding data and invoke a dependent sampling procedure. Therefore, theory behind adaptive design is much more complicated than that behind classic design. Validity and integrity have been strongly debated from statistical, operational, and regulatory perspectives during the past several years. However, despite different views, most scholars and practitioners believe that adaptive design could prove to be efficient tools for drug development if used properly. The issues of validity and integrity will be discussed in depth in Chapter 18.

Many interesting methods for adaptive design have been developed. Virtually all methods can be viewed as some combination of stagewise p-values. The stagewise p-values are obtained based on the subsample from each stage; therefore, they are mutually independent and uniformly distributed over [0, 1] under the null hypothesis. The first method uses the same stopping boundaries as a classic group sequential design (Pocock, 1977; O'Brien and Fleming, 1979), and allows stopping for early efficacy or futility. Lan and DeMets (1983) proposed the error spending method (ESM), in which the timing and number of analyses can be changed based on a prespecified error-spending function. ESM is derived from Brownian motion. The method has been extended to allow for sample-size re-estimation (SSR) (Cui, Hung, and Wang, 1999). It can be viewed as a fixed-weight method (i.e., using fixed weights for z-scores from the first and second stages re-

gardless of sample-size change). Lehmacher and Wassmer (1999) further degeneralized this weight method by using the inverse-normal method, in which the z-score is not necessarily taken from a normal endpoint, but from the inverse-normal function of stagewise p-values. Hence, the method can be used for any type of endpoint.

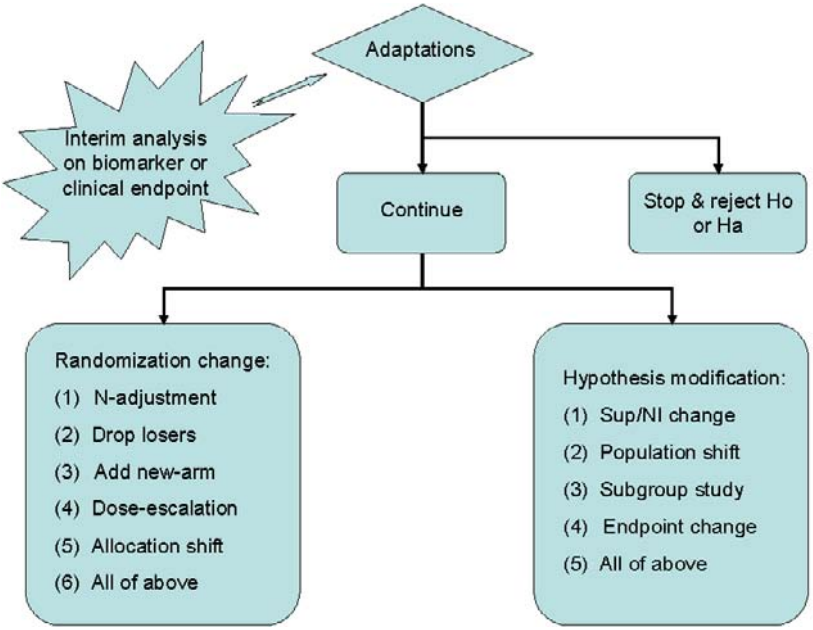


Figure 3.1: Various Adaptations

The second method is based on a direct combination of stagewise p-values. Bauer and Kohne (1994) use the Fisher combination (product) of stagewise p-values to derive the stopping boundaries. Chang (2006) used the sum of the stagewise p-values to construct a test statistic and derived a closed form for determination of stopping boundaries and p-value calculations as well as conditional power for trial monitoring.

The third method is based on the conditional error function. Proschan and Hunsberger (1995) developed an adaptive design method based the conditional error function for two-stage designs with Normal test statistics. Müller and Schäfer (2001) developed the conditional error method where the conditional error function is avoided and replaced with a conditional error that is calculated on fly. Instead of a two-stage design, Müller and Schäfer's method can be applied to a k-stage design and allows for many adaptations.

The fourth method is based on recursive algorithms such as Brannath-Posch-Bauer's recursive combination tests (Brannath, Posch and Bauer, 2002); Müller-Schäfer's decision-function method (Müller and Schäfer, 2004); and Chang's (2006) recursive two-stage adaptive design (RTAD). All four recursive methods are developed for k-stage designs allowing for general adaptations. RTAD is the simplest and most powerful method and the calculations of stopping boundary, conditional power, sample-size modification, p-values, and other operating characteristics can be performed manually without any difficulties. The major methods of adaptive designs in this book are presented in Figure 3.2

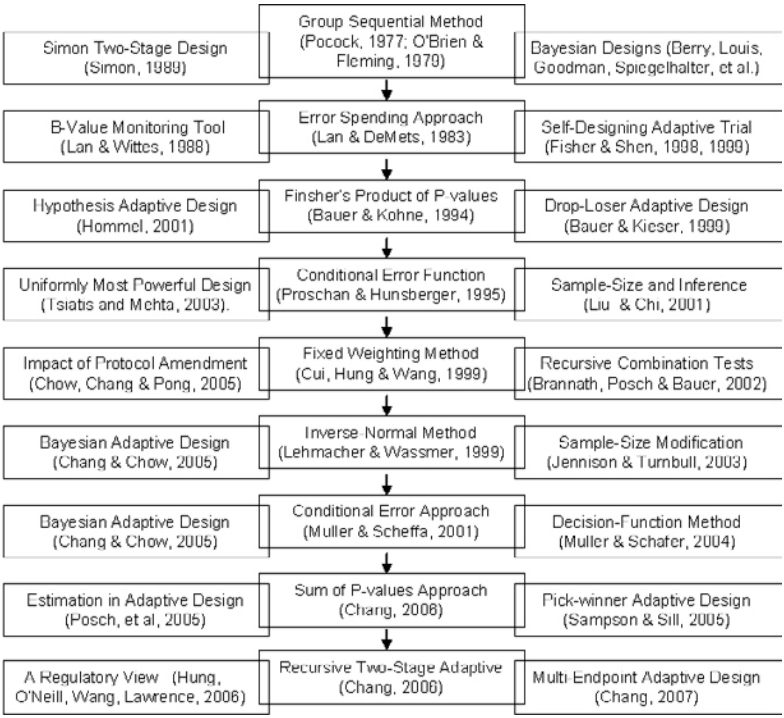


Figure 3.2: Selected Adaptive Design Methods from This Book

In the next several chapters we will cover these methods in detail, but now let's introduce the general framework for an adaptive design. Under this general framework, we can easily study the different methods, perform comparisons, and look into the relationships that exist among the different methods. This chapter will focus on three major issues: type-I error control, analysis including point and confidence interval estimations, and design evaluations.

This chapter might seem a bit too theoretical or abstract to readers who

are new to adaptive designs, but I hope you can read it adaptively, that is, just pay attention to the logic, ignoring the mathematical details. You should revisit this chapter from time to time after you read more in later chapters.

3.2 General Theory

There are four major components of adaptive designs in the frequentis paradigm: (1) type-I error rate or α - control: determination of stopping boundaries, (2) type-II error rate β: calculation of power or sample-size, (3) trial monitoring: calculation of conditional power or futility index, and (4) analysis after the completion of a trial: calculations of adjusted p-values, unbiased point estimates, and confidence intervals.

3.2.1 *Stopping Boundary*

Consider a clinical trial with K stages and at each stage a hypothesis test is performed followed by some actions that are dependent on the analysis results. Such actions can be early futility or efficacy stopping, sample-size re-estimation, modification of randomization, or other adaptations. The objective of the trial (e.g., testing the efficacy of the experimental drug) can be formulated using a global hypothesis test, which is the intersection of the individual hypothesis from the interim analyses.

$$H_o : H_{o1} \cap ... \cap H_{oK}, \tag{3.1}$$

where H_{ok} $(k = 1, ..., K)$ is the null hypothesis at the k^{th} interim analysis. Let's denote the sample-size per group for the subsample at the k^{th} stage as n_k. Note that the H_{ok} have some restrictions, that is, rejection of any H_{ok} $(k = 1, ..., K)$ will lead to the same clinical implication (e.g., drug is efficacious). Otherwise the global hypothesis can not be interpreted. In the rest of the chapter, H_{ok} testing will be based on subsamples from previous stages with the corresponding test statistic denoted as T_k which will be a combination of $p_i (i = 1, ..., k)$, where p_i is the p-value from the subsample obtained at the i^{th} stage. A one-sided test is always used in this book unless otherwise specified.

The stopping rules are given by

$$\begin{cases} \text{Stop for efficacy} & \text{if } T_k \leq \alpha_k\,, \\ \text{Stop for futility} & \text{if } T_k > \beta_k, \\ \text{Continue with adaptations} & \text{if } \alpha_k < T_k \leq \beta_k, \end{cases} \tag{3.2}$$

where $\alpha_k < \beta_k$ $(k = 1, ..., K-1)$, and $\alpha_K = \beta_K$. For convenience, α_k and β_k are called the efficacy and futility boundaries, respectively.

To reach the k^{th} stage, a trial has to pass the 1^{th} to $(k-1)^{th}$ stages. Therefore the c.d.f. of T_k is given by

$$\begin{aligned} \psi_k(t) &= \Pr(\alpha_1 < T_1 < \beta_1, ..., \alpha_{k-1} < T_{k-1} < \beta_{k-1}, T_k < t) \\ &= \int_{\alpha_1}^{\beta_1} ... \int_{\alpha_{k-1}}^{\beta_{k-1}} \int_{-\infty}^{t} f_{T_1 ... T_k}\, dt_k\, dt_{k-1} ... dt_1, \end{aligned} \tag{3.3}$$

where $f_{T_1 ... T_k}$ is the joint p.d.f. of $T_1, ...,$ and T_k.

Note that because the sequential adaptive designs control the overall alpha under a global null hypothesis (3.1), it is very important to properly form the hypothesis at each stage such that they are consistent and rejecting any of them will lead to the same clinical implication. This is particularly important when making hypothesis adaptations.

3.2.2 *Formula for Power and Adjusted P-value*

Definition 3.1: The p-value associated with a test is the smallest significance level α for which the null hypothesis is rejected (Robert, 1997, p.196).

Let $\psi_k(t)$ denote the probability that the test statistic T_k is equal or more extreme (smaller) than t. In other words, when H_0 (or $\delta = 0$) is true, $\psi_k(t)$ is the p-value if the trial stopped at the k^{th} stage. For convenience we call this p-value the conditional p-value and denote it as

$$p_c(t; k) = \psi_k(t|H_0). \tag{3.4}$$

The conditional p-value alone is not very informative because it does not measure the overall evidence against the null hypothesis. The conditional error rate (α spent) at the k^{th} stage is given by $\psi_k(\alpha_k|H_0)$; that is, the conditional error is

$$\pi_k = \psi_k(\alpha_k|H_0). \tag{3.5}$$

The conditional power of rejecting H_o at the k^{th} stage is given by

$$\varpi_k = \psi_k(\alpha_k|H_a). \tag{3.6}$$

At the efficacy stopping boundary, the conditional p-value $p_c(\alpha_k; k) = \psi_k(\alpha_k|H_0)$, irrespective of the choices of type-I error rate α.

When efficacy is claimed at a certain stage, the trial is stopped. Therefore, the type-I errors at different stages are mutually exclusive. Hence the experiment-wise type-I error rate can be written as

$$\alpha = \sum_{k=1}^{K} \pi_k. \tag{3.7}$$

Similarly, the power is given by

$$power = \sum_{k=1}^{K} \varpi_k. \tag{3.8}$$

Equation (3.5) is the key to determining the stopping boundaries adaptive designs as illustrated in the next several chapters. When $\Sigma_{i=1}^{k}\pi_i$ is reviewed as a function of information time or stage k, it is the so-called error-spending function.

There are several possible definitions of (adjusted) p-values. Here we are most interested in the so-called stagewise-ordering p-values (Jennison and Turnbull, 2000, p.180 and 356). Based on stagewise-ordering, extremeness is defined as $T_1 > T_2 > ... > T_K$, and within the same stage, extremeness is defined by their values. Therefore, the stagewise-ordering adjusted p-value is given by,

$$p(t; k) = \sum_{i=1}^{k-1} \pi_i + p_c(t; k). \tag{3.9}$$

An important characteristic of the adjusted p-value is that when the test statistic t is on stopping boundary a_k, p_k must be equal to alpha spent so far.

Note that the adjusted p-value is a measure of overall statistical strength against H_o. The later the H_o is rejected, the larger the adjusted p-value is, and the weaker the statistical evidence (against H_o) is. A late rejection leading to a larger p-value is reasonable because the alpha at earlier stages has been spent.

3.2.3 *Selection of Test Statistics*

Without losing generality, assume H_{ok} is a null hypothesis for the efficacy of the experimental drug, which can be written as

$$H_{ok} : \eta_{k1} \geq \eta_{k2} \quad \text{vs.} \quad H_{ak} : \eta_{k1} < \eta_{k2}, \tag{3.10}$$

where η_{k1} and η_{k2} are the treatment responses (mean, proportion, or survival) in the two comparison groups at the k^{th} stage.

It is desirable to chose T_k such that $f_{T_1 \ldots T_k}$ has a simple form. Notice that when $\eta_{k1} = \eta_{k2}$, the p-value p_k from the subsample at the k^{th} stage is uniformly distributed on [0,1] under H_o. This desirable property can be used to construct test statistics for adaptive designs.

There are many possible combinations of the p-values such as (1) linear combination (Chang, 2006)

$$T_k = \Sigma_{i=1}^{k} w_{ki} p_i, \ k = 1, ..., K, \tag{3.11}$$

(2) product of stagewise p-values (Fisher combination, Bauer and Kohne, 1994),

$$T_k = \prod_{i=1}^{k} p_i, \ k = 1, ..., K, \tag{3.12}$$

and (3) linear combination of inverse-normal stagewise p-values (Lehmacher and Wassmer, 1999, Cui, Hung, and Wang, 1999, Lan and DeMets, 1983)

$$T_k = \Sigma_{i=1}^{k} w_{ki} \Phi^{-1} (1 - p_i), \ k = 1, ..., K, \tag{3.13}$$

where weight $w_{ki} > 0$ can be constant or a function (ESM) of data from previous stages, and K is the number of analyses planned in the trial.

Note that p_k is the naive p-value from the subsample at the k^{th} stage, while $p_c(t; k)$ and $p(t; k)$ are conditional and adjusted p-values.

3.2.4 *Polymorphism*

After selecting the type of test statistic, we can determine the stopping boundaries α_k and β_k by using (3.3), (3.5), and (3.7) under the global null hypothesis (3.1). Once the stopping boundaries are determined, the power and sample-size under a particular H_a can be numerically calculated using (3.3), (3.6), and (3.8).

The polymorphism refers to the fact that the stopping boundaries (and other operating characteristics) can be constructed in many different ways, all with type-I error control.

After selecting the test statistic, we can choose one of the following approaches to fully determine the stopping boundaries:

(1) Choose certain types of functions for α_k and β_k. The advantage of using a stopping boundary function is that there are only limited parameters in the function to be determined. After the parameters are determined, the stopping boundaries are then fully determined using (3.3), (3.5), and (3.7), regardless of the number of stages. The commonly used boundaries are OB-F (O'Brien and Fleming, 1979), Pocock's (Pocock 1977), and Wang-Tsiatis' boundaries (Wang and Tsiatis, 1987).

(2) Choose certain forms of functions for π_k such that $\Sigma_{k=1}^{K}\pi_k = \alpha$. Traditionally, the cumulative quantity $\pi_k^* = \Sigma_{i=1}^{k}\ \pi_i$ is called the error-spending function, which can be either a function of stage k or the so-called information time based on sample-size fraction. After determining the function π_k or equivalently π_k^*, the stopping boundaries α_k,and β_k $(k = 1, ..., K)$ can be determined using (3.3), (3.5), and (3.7).

The so-called error-spending approach, which uses a predetermined error-spending function, allows for changing the number and timing of interim analyses. It is interesting to know that there is usually an equivalent stopping boundary function (at least implicitly) for any error-spending function (Lan and DeMets, 1983).

(3) Choose non-parametric stopping boundaries, i.e., no function is assumed, instead, use computer simulations to determine the stopping boundaries via a trial-error method. The non-parametric method does not allow for the changes to the number and timing of the interim analyses.

(4) Conditional error function method: One can rewrite the stagewise error rate for a two-stage design (see Chapter 8 for general multiple-stage designs) as

$$\pi_2 = \psi_2(\alpha_2|H_o) = \int_{\alpha_1}^{\beta_1} A(p_1)\, dp_1, \tag{3.14}$$

where $A(p_1)$ is called the conditional error function. For a given α_1 and β_1, by carefully selecting $A(p_1)$, the overall α control can be met (Proschan and Hunsberger, 1995). However, finding a good $A(p_1)$ isn't always easy. Therefore, the following method was developed.

(5) Conditional error method: similar to the conditional error function method, but in this method, for a given α_1 and β_1, $A(p_1)$ is calculated on-fly or in real-time, and only for the observed $\hat{p}_1$ under H_o. Adaptations

can be made under the condition that keep $A(p_1|H_o)$ unchanged.

Note that α_k and β_k are usually only functions of stage k or information time, but they can be functions of response data from previous stages, i.e., $\alpha_k = \alpha_k(t_1, ..., t_{k-1})$ and $\beta_k = \beta_k(t_1, ..., t_{k-1})$. In fact using variable transformation of the test statistic to another test statistic, the stopping boundaries often change from response-independent to response-dependent. Bauer and Kohne (1994) actually use a response-dependent boundary for the second stage α_2/p_1.

(6) Recursive two-stage design (Chang 2006) is a simple and powerful approach to a general N-stage design. It is considered an N-stage adaptive design, and is a composite of many overlapped two-stage designs that use the conditional error principal to derive the closed forms for the N-stage design (see Chapter 8 for details).

If you feel you have had enough math, you can skip to the next chapter and come back to this chapter after Chapter 7.

3.2.5 *Adjusted Point Estimates*

Estimation problems deserve a bit of philosophical discussion before we proceed to how to calculate them. We have focused our discussion within the frequentist paradigm, which is constructed fundamentally on the concept of repeated experiments. There are at least three types of unbiased point estimates, corresponding to four different sample spaces: (1) Unconditional point estimate (UE); the corresponding sample space consists of all possible results from a repeated experiments with a given design. Usually only the sponsor can see these results; (2) Conditional estimate (CE) that is based on the positive or statistically significant results; the corresponding sample space consists of all results with statistical significance. Regulatory authorities and patients usually only see this set of results; (3) Stagewise estimate (SE), which is based on trial stopping at each stage; the corresponding sample space is all possible results from repeated experiments when they stop at a given stage k.

Theoretically, the sponsor (pharmaceutical company) can see all POSSIBLE results from a trial (equivalent to all results from repeated experiments with a given design); sponsors usually only submit positive trial results to regulatory agencies, and the agencies weigh the benefit-risk ratio and select a subset of the positive results for approval for marketing. For a classic, single-stage design, what sponsors see is the unconditional estimate. What the regulatory agencies and patients see are roughly the conditional estimates.

Conditional Estimate

Let's take a hypothesis testing two group means as an example. Here we will discuss CE (δ_c) and UE (δ) for both classic and adaptive designs. For a normal response, the CE is the mean under the condition that the null hypothesis of no treatment effect is rejected. It can be derived that the relative bias of the conditional mean for a classic design with two independent groups is given by

$$\frac{\delta - \delta_c}{\delta} = \frac{1}{1-\beta}\frac{\sigma}{\delta\sqrt{\pi n}}\exp\left(-\frac{1}{2}z_{1-\beta}^2\right), \tag{3.15}$$

where β is the type-II error rate, σ is the standard deviation, and n is the sample-size per group. It is true that what we submit to the regulatory reviewers is a conditional mean that is biased. For $\beta = 0.2$, there is about a 12% bias for a classic design (see Table 3.1). Whether a conditional or unconditional mean is submitted to regulatory authorities, the approval will be based on the conditional mean. Therefore, what patients see is the most biased mean. Statisticians are often faced with the question of whether to report the conditional or unconditional mean. Should the conditional mean be adjusted because it is reported to patients and is biased for both classic and adaptive designs?

Table 3.1: Conditional and Unconditional Means

Design	True mean difference	Unconditional mean difference	Conditional mean difference
Classic	1	1	1.12
Adaptive	1	1.05	1.25

Note: Standard deviation = 2.5. $N_{max} = N_{fix} = 100$/group

Which mean should be used under which condition? If the conditional mean is the most important because it is what sponsors show the FDA and patients, then it should be adjusted regardless of classic or adaptive design because it is biased in both designs. Because the conditional mean (CM) is biased for both classic and adaptive designs, there is no reason to adjust it for an adaptive design but not for a classic design.

Unconditional Estimate

If the unconditional mean is the most important, then it should be adjusted for an adaptive design, but not for a classic design. A general method for obtaining an unbiased point estimate is described as follows:

Let δ_B be a biased estimate for an adaptive design, and δ be the true value for the parameter of interest. The bias can be expressed as a function

of δ:

$$\xi(\delta) = \delta - \bar{\delta}_B, \tag{3.16}$$

where $\bar{\delta}_B$ is the expectation of δ_B.

$$\bar{\delta}_B = \delta - \xi(\delta). \tag{3.17}$$

From (3.17) we obtain

$$\delta = \eta^{-1}(\bar{\delta}_B), \tag{3.18}$$

where $\eta^{-1}(\delta)$ is the inverse function of $\eta(\delta) = \delta - \xi(\delta) = (I - \xi)(\delta)$, $I =$ identity mapping. An unbiased estimate can be given by

$$\delta_u = \delta_B + \xi(\delta) = \delta_B + \xi\left((I - \xi)^{-1}\left(\bar{\delta}_B\right)\right). \tag{3.19}$$

The challenge is that we don't know δ and $\xi(\cdot)$. However, we can use linear approximation to $\xi(\delta)$ to solve the problem.

Assume

$$\xi(\delta) = c_0 + c_1\delta, \tag{3.20}$$

where c_i $(i = 0, 1)$ are constants. Substituting (3.20) into (3.17) and solving for δ, we can obtain

$$\delta = \frac{c_0 + \bar{\delta}_B}{1 - c_1}. \tag{3.21}$$

Because $c_i (i = 0, 1)$ is a constant, we can immediately obtain an unbiased estimator from (3.21):

$$\delta_u = \frac{c_0 + \delta_B}{1 - c_1}. \tag{3.22}$$

The monotonic relationship $\xi(\delta) = c_0 + c_1\delta$ usually holds at least in a small range of $\delta \in (\delta - \varepsilon, \delta + \varepsilon)$.

By trying several (at least 2) $\delta = \delta_m$ around δ_B, and using simulation to calculate the bias $\xi(\delta_m) = \delta_m - \bar{\delta}_{B_m}$ for each δ_m, we get

$$\xi(\delta_m) = c_0 + c_1\delta_m. \tag{3.23}$$

We can solve for (if only $m = 2$) or estimate c_0 and c_1 from (3.23).

The reasons to choose the values of δ_m near the best guessed value of δ_B are obvious. If δ_B is near δ, then (3.23) works well; if δ_B is far away

from the true δ, adjusting is not important anyway. If we name (3.23) the first-order bias adjustment, then the zero-order bias adjustment is a degeneralized case when letting $c_1 = 0$ in (3.23).

Stagewise Estimates

The research papers on adjusted stagewise estimates are Lawrence and Hung, 2003; Wassmer 2005; Posch et al., 2006; Brannath, Konig, and Bauer, 2006, among others. The sample space for the stagewise estimate at the k^{th} stage consists of all possible outcomes when the trial is stopped at the k^{th} stage.

Consider the following estimate:

$$\delta_{sw} = \sum_{i=1}^{k} w_{ki}\delta_i, \text{ if trial stops at the } k^{th} \text{ stage,} \tag{3.24}$$

where δ_i is the stagewise unbiased estimate of treatment difference δ based on the subsample from the i^{th} stage; k is the stage where the trial stops; and w_{ki} is a constant weight that usually, but not necessarily satisfies $\Sigma_{i=1}^{k} w_{ki} = 1$.

If the trial design does not allow for early stopping, e.g., an interim analysis (IA) for sample-size adjustment only, then (3.24) is an unbiased estimator of δ. This is because

$$E(\delta_{sw}) = \sum_{i=1}^{k} w_{ki} E(\delta_i) = \sum_{i=1}^{k} w_{ki}\delta = \delta \sum_{i=1}^{k} w_{ki} = \delta. \tag{3.25}$$

However, most adaptive clinical trials do allow for early stopping and δ_{sw} from (3.24) is biased in general. A simple solution to get unconditionally unbiased estimates is to add a few subjects (at least two) even if the trial has been stopped early. These few subjects will not be used for p-value calculation, only to get an unbiased estimation.

3.2.6 *Derivation of Confidence Intervals*

Consider the null hypothesis:

$$H : \delta = \delta_0. \tag{3.26}$$

For $\delta_0 = 0$, we use the test statistic T. In general, we use the test statistic

$$\tilde{T} = T - T_0(\delta_0), \tag{3.27}$$

where the function $T_0(\delta_0)$ is the expectation of T under H, and $T_0(0) = 0$.

A 100(1-α)% confidence interval consists of all the δ_0 such that the null hypothesis (3.26) would not be rejected, given the observed value $\hat{T}$ of the test statistic $\tilde{T}$. Therefore, the upper and lower bounds of this confidence interval are found by equating $\tilde{T} = \alpha_k$ at the k^{th} stage (Assume that $\tilde{T}$ under $\delta = \delta_0$ and T under $\delta = 0$ have the same distribution; therefore the same stopping boundary α_k can be used.). This leads to

$$T - T_0(\delta_0) = \pm\alpha_k. \tag{3.28}$$

Equation (3.28) can be solved for δ_0 to obtain the confidence limits:

$$\delta_0 = T_0^{-1}\left(\hat{T} \mp \alpha_k\right). \tag{3.29}$$

If the stagewise-ordering adjusted p-value is used, then (3.29) is also numerically equal to the $(1-\Sigma_{i=1}^{k}\pi_i)$% confidence limits if the trial is stopped at the k^{th} stage.

For an adaptive design, if the test statistic at the k^{th} stage is given by

$$T = \sum_{i=1}^{k} w_{ki}\frac{\delta_i}{\sigma}\sqrt{\frac{n_i}{2}}, \tag{3.30}$$

where $\Sigma_{i=1}^{K} w_{ki}^2 = 1$, then we have

$$T_o(\delta_0) = \sum_{i=1}^{k} w_{ki}\frac{\delta_0}{\sigma}\sqrt{\frac{n_i}{2}}. \tag{3.31}$$

For symmetrical stopping boundaries (i.e., two-sided α, one efficacy boundary and one futility boundary, they are symmetrical), the lower and upper limits of a $(1 - \Sigma_{i=1}^{k}\pi_i)$% confidence interval at the k^{th} stage are given by

$$\delta_0 = \frac{\sum_{i=1}^{k} w_{ki}\frac{\delta_i}{\sigma}\sqrt{\frac{n_i}{2}} \mp \alpha_k}{\sum_{i=1}^{k} w_{ki}\frac{1}{\sigma}\sqrt{\frac{n_i}{2}}}. \tag{3.32}$$

Also, because of symmetry, the point estimate is given by

$$\delta_0 = \frac{\sum_{i=1}^{k} w_{ki}\delta_i\sqrt{\frac{n_i}{2}}}{\sum_{i=1}^{k} w_{ki}\sqrt{\frac{n_i}{2}}}. \tag{3.33}$$

For a one-sided confidence limit, one of the limits is set at infinity, and (3.32) and (3.33) hold approximately. Note that in calculation, we can replace σ by $\hat{\sigma}$. (3.32) can be viewed as the $(1-\alpha)\%$ repeated confidence interval (RCI).

3.3 Design Evaluation - Operating Characteristics

3.3.1 *Stopping Probabilities*

The stopping probability at each stage is an important property of an adaptive design, because it provides the time-to-market and the associated probability of success. It also provides information on the cost (sample-size) of the trial and the associated probability. In fact, the stopping probabilities are used to calculate the expected samples that present the average cost or efficiency of the trial design and the duration of the trial.

There are two types of stopping probabilities: unconditional probability of stopping to claim efficacy (reject H_o) and unconditional probability of futility (accept H_o). The former refers to the efficacy stopping probability (ESP), and the latter refers to the futility stopping probability (FSP). From (3.3), it is obvious that the ESP at the k^{th} stage is given by

$$ESP_k = \psi_k(\alpha_k) \tag{3.34}$$

and the FSP at the k^{th} stage is given by

$$FSP_k = 1 - \psi_k(\beta_k). \tag{3.35}$$

3.3.2 *Expected Duration of an Adaptive Trial*

The stopping probabilities can be used to calculate the expected trial duration, which is definitely an important feature of an adaptive design. The conditionally (on the efficacy claim) expected trial duration is given by

$$\bar{t}_e = \sum_{k=1}^{K} ESP_k \, t_k, \tag{3.36}$$

where t_k is the time from the first-patient-in to the k^{th} interim analysis.

The conditionally (on the futility claim) expected trial duration is given by

$$\bar{t}_f = \sum_{k=1}^{K} FSP_k \, t_k. \tag{3.37}$$

The unconditionally expected trial duration is given by

$$\bar{t} = \sum_{k=1}^{K} \left(ESP_k + FSP_k\right) t_k. \tag{3.38}$$

3.3.3 *Expected Sample Sizes*

The expected sample-size is a commonly used measure of the efficiency (cost and timing of the trial) of the design. The expected sample-size is a function of the treatment difference and its variability, which are unknowns. Therefore, expected sample-size is really based on hypothetical values of the parameters. For this reason, it is beneficial and important to calculate the expected sample-size under various critical or possible values of the parameters. The total expected sample-size per group can be expressed as

$$N_{\text{exp}} = \sum_{k=1}^{K} n_k \left(ESP_k + FSP_k\right) = \sum_{k=1}^{K} n_k \left(1 + \psi_k(\alpha_k) - \psi_k(\beta_k)\right). \tag{3.39}$$

It can also be written as

$$N_{\text{exp}} = N_{\max} - \sum_{k=1}^{K} n_k \left(\psi_k(\beta_k) - \psi_k(\alpha_k)\right), \tag{3.40}$$

where $N_{\max} = \sum_{k=1}^{K} n_k$ is the maximum sample-size per group.

3.3.4 *Conditional Power and Futility Index*

The conditional power is the conditional probability of rejecting the null hypothesis during the rest of the trial based on the observed interim data. The conditional power is commonly used for monitoring an ongoing trial. Similar to the ESP and FSP, conditional power is dependent on the population parameters or treatment effect and its variability. The conditional power at the k^{th} stage is the sum of the probability of rejecting the null hypothesis at stage $k+1$ to K (K does not have to be predetermined), given the observed data from stages 1 through k.

$$cP_k = \sum_{j=k+1}^{K} \Pr\left(\cap_{i=k+1}^{j-1}\left(a_i < T_i < \beta_i\right) \cap T_j \leq \alpha_j | \cap_{i=1}^{k} T_i = t_i\right), \quad (3.41)$$

where t_i is the observed test statistic T_i at the i^{th} stage. For a two-stage design, the conditional power can be expressed as

$$cP_1 = \Pr\left(T_2 \leq \alpha_2 | t_1\right). \quad (3.42)$$

The futility index is defined as the conditional probability of accepting the null hypothesis:

$$FI_k = 1 - cP_k. \quad (3.43)$$

3.3.5 *Utility and Decision Theory*

It is important to realize that the choice of a design should not be based on power only. In fact power may not be a good criterion for evaluating a design, especially when it comes to adaptive designs. In many situations, time is more important than power. Also, a design with high power to detect a small and clinically irrelevant difference is not desirable.

Decision theory is a body of knowledge that assists a decision maker in choosing among a set of alternatives in light of their possible consequences. Decision theory is based on the concept of utility or, equivalently, the loss function. The decision maker's preferences for the mutually exclusive consequences of an alternative are described by a utility function that permits calculation of the expected utility for each alternative using Bayesian theory. The alternative with the highest expected utility is considered the most preferable. The Bayesian decision theory can be illustrated in Figure 3.3.

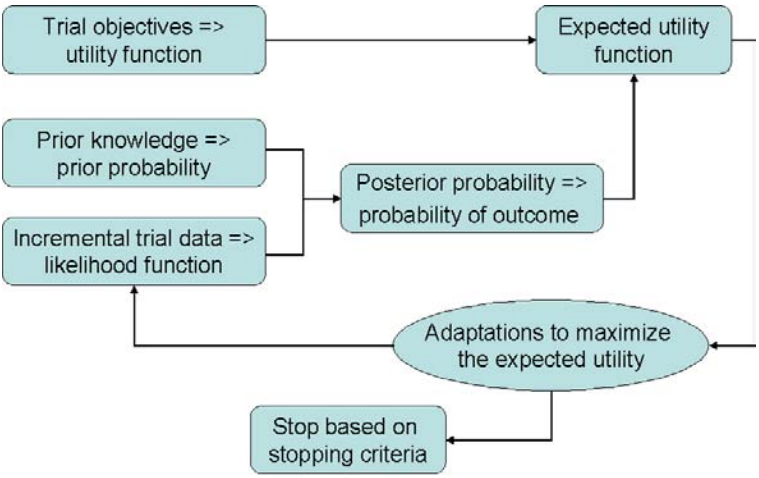


Figure 3.3: Bayesian Decision Approach

For an adaptive trial with sample-size re-estimation, the Bayesian decision theory can be briefly stated as follows:

Define the utility $U\left(\hat{\delta}, n\right)$, the prior distribution $\pi(\delta)$. Denote the posterior by $\pi\left(\delta|\hat{\delta}_1\right)$ and the interim observed treatment difference by $\hat{\delta}_1$.

The expected utility at the design stage is given by

$$EU(n) = \int\int U\left(\hat{\delta}, n\right) \pi(\delta) \Pr\left(\hat{\delta}|\delta\right) d\delta \, d\hat{\delta}$$

The expected utility at the interim analysis is given by

$$EU(n) = \int\int U\left(\hat{\delta}, n\right) \pi\left(\delta|\hat{\delta}_1\right) \Pr\left(\hat{\delta}|\delta\right) d\delta \, d\hat{\delta}$$

The Bayesian decision approach is to determine an action (i.e., sample-size n) that maximizes the expected utility under certain constraints:

$$\frac{\partial EU(n)}{\partial n} = 0.$$

Decision theory can be viewed as one-person game theory, involving a game with a single player against nature. This refers to a situation where the result of a decision depends on the action of another player (nature). For example, if the decision is to carry an umbrella or not, the result (get wet or not) depends on what action nature takes. An important feature of this model is that the returns only affect the decision maker, not nature. However, in game theory, both players have an interest in the outcome.

Note that we can treat our problems using utility theory because a decision maker's action does not materially affect nature. In the modern drug development age, pharmaceutical companies face many competitors. Therefore, strictly speaking, game theory is more applicable than utility theory in this competing-cooperative environment. Fortunately, utility theory can take the competition into consideration by constructing a utility that includes the result of the competitor's action (if it is relatively static), although this approach is not as good as game theory. We will discuss later in detail how to generate optimal designs using Bayesian decision theory in Chapter 16.

3.4 Summary

In this chapter, we have provided a uniform formulation for adaptive designs and the polymorphism, i.e., how the uniform formulation can be used to develop various adaptive design methods. In the next several chapters, we will illustrate this in detail, derive different methods, and apply them to different trials. We have also discussed the estimation issues and the general methods. In Chapter 8, we will discuss estimation again in great detail with trial examples. Evaluation of trial designs is obviously important. We have reviewed many operating characteristics of adaptive designs. Keep in mind that we should not misunderstand power to be the sole criterion when judging a trial design. Instead, think of drug development as an integrated process – decision or game theory is the ultimate tool for trial evaluation.

Problem

3.1 Prove that the bias of conditional estimate can be expressed by (3.15) for classic design (assume $E(\bar{x}) = \frac{\delta}{\sigma\sqrt{n}}$, where σ = standard deviation for the mean difference).

Chapter 4

Method with Direct Combination of P-values

In this chapter, we will use the theory developed in the previous chapter with several different test statistics based on direct combination of stagewise p-values, which includes: (1) method based on individual p-values (MIP), (2) method based on the sum of p-values (MSP), and (3) method based on the product of p-values (MPP). We will focus on two-stage designs and derive the closed forms for determination of stopping boundaries and adjusted p-values. Many different examples are presented, in which power and sample-size calculations are based on computer simulations. The methods are very general, meaning that they can be applied to broad adaptations. However, the examples provided will focus on classic group sequential designs and sample-size re-estimation (SSR).

4.1 Method Based on Individual P-values

This method refers to MIP, in which the test statistic is defined as

$$T_k = p_k, \tag{4.1}$$

where p_k is the stagewise p-value from the k^{th} stage subsample.

Using (3.3), (3.5), and (3.7), a level-α test requires:

$$\alpha = \sum_{k=1}^{K} \alpha_k \prod_{i=1}^{k-1} (\beta_i - \alpha_i)\,. \tag{4.2}$$

When the upper bound exceeds the lower bound in (4.2), define $\prod_{i=1}^{0}(\cdot) = 1$.

Using (4.2), the stopping boundary (α_i, β_i) can be determined. For a two-stage design, (4.2) becomes

$$\alpha = \alpha_1 + \alpha_2(\beta_1 - \alpha_1). \tag{4.3}$$

For convenience, examples of stopping boundaries from (4.2) are tabulated for a one-sided $\alpha = 0.025$ (Table 4.1).

Table 4.1: Stopping Boundaries α_2 with MIP

	α_1	0.000	0.0025	0.005	0.010	0.015	0.020
β_1							
0.15		0.1667	0.1525	0.1379	0.1071	0.0741	0.0385
0.20		0.1250	0.1139	0.1026	0.0789	0.0541	0.0278
0.25		0.1000	0.0909	0.0816	0.0625	0.0426	0.0217
0.30	α_2	0.0833	0.0756	0.0678	0.0517	0.0351	0.0179
0.35		0.0714	0.0647	0.0580	0.0441	0.0299	0.0152
0.50		0.0500	0.0452	0.0404	0.0306	0.0206	0.0104
1.00		0.0250	0.0226	0.0201	0.0152	0.0102	0.0051

Note: One-sided $\alpha = 0.025$.

The stagewise-ordering p-value defined by (3.9) is given by

$$p(t;k) = \begin{cases} t, & k = 1, \\ \alpha_1 + t(\beta_1 - \alpha_1) & k = 2\ . \end{cases} \tag{4.4}$$

MIP is useful in the sense that it is very simple and can serve as the "baseline" for comparing different methods. MIP does not use combined data from different stages, while most other adaptive designs do.

SAS Macro 4.1 has been implemented for simulating two-arm adaptive trials with a binary endpoint and allowing for sample-size re-estimation. The test statistic can be based on individual stagewise p-values, or the sum or product of the stagewise p-values (details provided later in this chapter). The SAS variables are defined as follows: **Px** and **Py** are true proportions of response in the groups x and y, respectively; **DuHa** = the estimate for the true treatment difference under the alternative H_a; **N** = sample-size per group; **alpha1** = early efficacy stopping boundary (one-sided); **beta1** = early futility stopping boundary (one-sided); and **alpha2** = final efficacy stopping boundary (one-sided). The null hypothesis test is H_o: δ+**NId** < 0, where δ = **Py** - **Px** is the treatment difference and **NId** = noninferiority margin (NId ≤ 0 for superiority and NId > 0 for noninferiority test). **nSims** = the number of simulation runs, **alpha** = one-sided overall type-I error rate, and **beta** = type-II error rate. **nAdj** = "N" for the

case without sample-size re-estimation and **nAdj** = "Y" for the case with sample-size adjustment, **Nmax** = maximum sample-size allowed, **N0** = the initial sample-size at the final analysis, **nInterim** = sample-size for the interim analysis, **a** = the parameter in (4.17) for the sample-size adjustment, **FSP** = futility stopping probability, **ESP** = efficacy stopping probability, **AveN** = average sample-size, **Power** = power of the hypothesis testing, **nClassic** = sample-size for the corresponding classic design, and **Model** = "ind", "sum", or "prd" for the methods, MIP, MSP, and MPP, respectively.

››SAS Macro 4.1: Two-Stage Adaptive Design with Binary Endpoint››

```
%Macro DCSPbinary(nSims=1000000, Model="sum", alpha=0.025,
                  beta=0.2, NId=0, Px=0.2, Py=0.4, DuHa=0.2,
                  nAdj="N", Nmax=100, N0=100, nInterim=50, a=2,
                  alpha1=0.01, beta1=0.15, alpha2=0.1871);
Data DCSPbinary; Keep Model FSP ESP AveN Power nClassic;
 seedx=2534; seedy=6762; Model=&model; NId=&NId;
 Nmax=&Nmax; N1=&nInterim; Px=&Px; Py=&Py;
 eSize=abs((&DuHa+NId)/((Px*(1-Px)+Py*(1-Py))/2)**0.5);
 nClassic=Round(2*((probit(1-&alpha)+probit(1-&beta))/eSize)**2);
 FSP=0; ESP=0; AveN=0; Power=0;
 Do isim=1 To &nSims;
 nFinal=N1;
    Px1=Ranbin(seedx,N1,px)/N1;
    Py1=Ranbin(seedy,N1,py)/N1;
    sigma=((Px1*(1-Px1)+Py1*(1-Py1))/2)**0.5;
    T1 = (Py1-Px1+NId)*Sqrt(N1/2)/sigma;
    p1=1-ProbNorm(T1);
    If p1>&beta1 Then FSP=FSP+1/&nSims;
    If p1<=&alpha1 Then Do;
        Power=Power+1/&nSims; ESP=ESP+1/&nSims;
    End;
    If p1>&alpha1 and p1<=&beta1 Then Do;
        eRatio=abs(&DuHa/(abs(Py1-Px1)+0.0000001));
        nFinal=Min(&Nmax,Max(&N0,eRatio**&a*&N0));
        If &nAdj="N" then nFinal=&Nmax;
        If nFinal>N1 Then Do;
            N2=nFinal-N1;
            Px2=Ranbin(seedx,N2,px)/N2;
            Py2=Ranbin(seedy,N2,py)/N2;
            sigma=((Px2*(1-Px2)+Py2*(1-Py2))/2)**0.5;
```

```
                T2 = (Py2-Px2+NId)*Sqrt(N2/2)/sigma;
                p2=1-ProbNorm(T2);
                If Model="ind" Then TS2=p2;
                If Model="sum" Then TS2=p1+p2;
                If Model="prd" Then TS2=p1*p2;
                If .<TS2<=&alpha2 then Power=Power+1/&nSims;
            End;
        End;
        AveN=Aven+nFinal/&nSims;
    End;
    Output;
Run;
Proc Print Data=DCSPbinary; Run;
%Mend DCSPbinary;
```

‹‹**SAS**‹‹

Example 4.1 Adaptive Design for Acute Ischemic Stroke Trial

A phase-III trial is to be designed for patients with acute ischemic stroke of recent onset. The composite endpoint (death and MI) is the primary endpoint, and the event rate is 14% for the control group and 12% for the test group. Based on a large sample assumption, the sample-size for a classic design is 5937 per group, which has 90% power to detect the difference at the one-sided alpha = 0.025. Using MIP, an interim analysis is planed based on a response assessment of 50% of the patients. We use SAS macro 4.1 to design the trial as follows:

(1) Choose stopping boundaries at the first stage: $\alpha_1 = 0.01$, $\beta_1 = 0.25$; then from Table 4.1, we obtain $\alpha_2 = 0.0625$.

(2) Check the stopping boundary to make sure that the familywise error is controlled. We run simulations under the null hypothesis (14% event rate for both groups) using the following SAS code:

```
››SAS››
%DCSPbinary( Model="ind", alpha=0.025, beta=0.1, Px=0.14,
Py=0.14, DuHa=0.02, nAdj="N", Nmax=7000, nInterim=3500,
alpha1=0.01, beta1=0.25, alpha2=0.0625);
```

‹‹**SAS**‹‹

The simulated familywise error rate is $\alpha = 0.0252$; therefore the stopping boundaries are confirmed.

(3) Calculate power or sample-size under the alternative hypothesis (14% and 12% event rates for the control and the test groups, respectively) using the following SAS code:

```
>>SAS>>
%DCSPbinary(Model="ind", alpha=0.025, beta=0.1, Px=0.12, Py=0.14,
DuHa=0.02, nAdj="N", Nmax=7000, nInterim=3500, alpha1=0.01,
beta1=0.25, alpha2=0.0625);
<<SAS<<
```

(4) Perform sensitivity analyses (under condition H_s). Because the treatment difference is unknown, it is desirable to perform simulations under different assumptions about treatment difference, e.g., treatment difference = 0.015 (14% versus 12.5%). For the sensitivity analysis, we simply use the following SAS macro call:

```
>>SAS>>
%DCSPbinary(Model="ind", alpha=0.025, beta=0.1, Px=0.125, Py=0.14,
DuHa=0.015, nAdj="N", Nmax=7000, nInterim=3500, alpha1=0.01,
beta1=0.25, alpha2=0.0625);
<<SAS<<
```

We now can summarize the simulation outputs of the three scenarios in Table 4.2.

Table 4.2: Operating Characteristics of a GSD with MIP

Simulation condition	FSP	ESP	$\bar{N}$	N_{max}	Power (alpha)
H_o	0.750	0.010	4341	7000	(0.025)
H_a	0.035	0.564	4905	7000	0.897
H_s	0.121	0.317	5468	7000	0.668

From Table 4.2, we can see that the design has a smaller expected sample-size ($\bar{N}$) under H_o and H_a (4341, 4905) than the classic design (5937). However, the group sequential design has a larger maximum sample-size (7000) than the classic design (5937). The early futility stopping probability (FSP) and early efficacy stopping probability (ESP) are also shown in Table 4.2. The sensitivity analysis shows a large power loss when treatment difference is lower than expected, i.e., the group sequential design does not protect the power when the initial effect size is overestimated. To protect power, we can use the sample-size re-estimation method, which will be discussed later in this chapter.

Note that the MIP design is different from the sequence of two separate trials because MIP can use the early futility boundary in constructing a later stopping boundary. We will discuss this issue in great detail later.

Now let's calculate adjusted p-values. Assume that the trial is finished, with the stagewise p-value $p_1 = 0.012$ (which is larger than α_1 =0.01 and

not significant; therefore the trial continues to the second stage) and $p_2 = 0.055 < \alpha_2 = 0.0625$. Therefore, the null hypothesis is rejected, and the test drug is significantly better than the control.

The stagewise-ordering p-value can be calculated from (4.4) using $t = p_2$:

$$p = \alpha_1 + p_2(\beta_1 - \alpha_1) = 0.01 + 0.055(0.25 - 0.01) = 0.0232.$$

4.2 Method Based on the Sum of P-values

Chang (2006) proposed an adaptive design method, in which the test statistic is defined as the sum of the stagewise p-values. This method is referred to as MSP. The test statistic is defined as

$$T_k = \Sigma_{i=1}^{k} p_i, \ k = 1, ..., K. \tag{4.5}$$

For two-stage designs, the α spent at stage 1 and stage 2 is given by

$$\pi_1 = \int_0^{\alpha_1} dt_1 = \alpha_1 \tag{4.6}$$

and

$$\pi_2 = \begin{cases} \int_{\alpha_1}^{\beta_1} \int_{t_1}^{\alpha_2} dt_2 dt_1 \text{ for } \beta_1 \leq \alpha_2, \\ \int_{\alpha_1}^{\alpha_2} \int_{t_1}^{\alpha_2} dt_2 dt_1 \text{ for } \beta_1 > \alpha_2, \end{cases} \tag{4.7}$$

respectively.

Carrying out the integrations in (4.7) and substituting the results and (4.6) into (3.7), we immediately obtain the following formulation for determining the stopping boundaries:

$$\alpha = \begin{cases} \alpha_1 + \alpha_2(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2) & \text{for } \beta_1 < \alpha_2, \\ \alpha_1 + \frac{1}{2}(\alpha_2 - \alpha_1)^2 & \text{for } \beta_1 \geq \alpha_2. \end{cases} \tag{4.8}$$

To calculate the stopping boundaries for given α, α_1, and β_1, solve (4.8) for α_2.Various stopping boundaries can be chosen from (4.8). See Table 4.3 for examples of the stopping boundaries.

Table 4.3: Stopping Boundaries α_2 with MSP

	α_1	0.000	0.0025	0.005	0.010	0.015	0.020
β_1							
0.05		0.5250	0.4999	0.4719	0.4050	0.3182	0.2017
0.10		0.3000	0.2820	0.2630	0.2217	0.1751	0.1225
0.15	α_2	0.2417	0.2288	0.2154	0.187 1	0.1566	0.1200
0.20		0.2250	0.2152	0.2051	0.1832	0.1564	0.1200
>0.25		0.2236	0.2146	0.2050	0.1832	0.1564	0.1200

Note: One-sided $\alpha = 0.025$.

The stagewise-ordering p-value can be obtained by replacing α_1 with t in (4.6) if the trial stops at stage 1 and by replacing α_2 with t in (4.8) if the trial stops at stage 2. That is

$$p(t;k) = \begin{cases} t, & k = 1, \\ \alpha_1 + t(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2), & k = 2 \text{ and } \beta_1 < \alpha_2, \\ \alpha_1 + \frac{1}{2}(t - \alpha_1)^2, & k = 2 \text{ and } \beta_1 \geq \alpha_2. \end{cases} \tag{4.9}$$

It is interesting to know that when $p_1 > \alpha_2$, there is no point in continuing the trial because $p_1 + p_2 > p_1 > \alpha_2$, and futility should be claimed. Therefore, statistically it is always good idea to choose $\beta_1 < \alpha_2$. However, because the non-binding futility rule is adopted currently by the regulatory bodies, it is better to use the stopping boundaries with $\beta_1 = \alpha_2$.

The SAS Macro 4.2 is implemented for simulating two-arm adaptive designs with a Normal endpoint. The adaptive method can be based on individual stagewise p-values, or the sum or product of the stagewise p-values (see details later in this chapter). The SAS variables are defined as follows: **ux** and **uy** are true treatment means in the x and y groups, respectively. **DuHa** = the estimate for the true treatment difference under the alternative H_a, and **N** = sample-size per group. **alpha1** = early efficacy stopping boundary (one-sided), **beta1** = early futility stopping boundary (one-sided), and **alpha2** = final efficacy stopping boundary (one-sided). The null hypothesis test is H_o: δ+**NId** < 0, where δ = **uy** - **ux** is the treatment difference, and **NId** = noninferiority margin. **nSims** = the number of simulation runs, **alpha** = one-sided overall type-I error rate, and **beta** = type-II error rate. **nAdj** = "N" for the case without sample-size re-estimation and **nAdj** = "Y" for the case with sample-size adjustment, **Nmax** = maximum sample-size allowed; **N0** = the initial sample-size at the final analysis; **nInterim** = sample-size for the interim analysis; **a** = the parameter in (4.17) for the sample-size adjustment; **FSP** = futility stopping probability; **ESP** = efficacy stopping probability; **AveN** = average sample-size; **Power** = power of the hypothesis testing; **nClassic** = sample-size for

the corresponding classic design; and **Model** = "ind", "sum", or "prd" for the methods MIP, MSP, and MPP, respectively.

››SAS Macro 4.2: Two-Stage Adaptive Design with Normal Endpoint››

```
%Macro DCSPnormal(nSims=1000000, Model="sum", alpha=0.025,
        beta=0.2, sigma=2, NId=0, ux=0, uy=1, nInterim=50,
        Nmax=100, N0=100, DuHa=1, nAdj="Y", a=2,
        alpha1=0.01, beta1=0.15, alpha2=0.1871);
Data DCSPnormal; Keep Model FSP ESP AveN Power nClassic;
 seedx=1736; seedy=6214; alpha=&alpha; NId=&NId; Nmax=&Nmax;
 ux=&ux; uy=&uy; sigma=&sigma; model=&Model; N1=&nInterim;
 eSize=abs(&DuHa+NId)/sigma;
 nClassic=round(2*((probit(1-alpha)+probit(1-&beta))/eSize)**2);
 FSP=0; ESP=0; AveN=0; Power=0;
 Do isim=1 To &nSims;
 nFinal=N1;
     ux1 = Rannor(seedx)*sigma/Sqrt(N1)+ux;
     uy1 = Rannor(seedy)*sigma/Sqrt(N1)+uy;
     T1 = (uy1-ux1+NId)*Sqrt(N1)/2**0.5/sigma;
     p1=1-ProbNorm(T1);
     If p1>&beta1 then FSP=FSP+1/&nSims;
     If p1<=&alpha1 then do;
         Power=Power+1/&nSims; ESP=ESP+1/&nSims;
     End;
     If p1>&alpha1 and p1<=&beta1 Then Do;
         eRatio = abs(&DuHa/(abs(uy1-ux1)+0.0000001));
         nFinal = min(&Nmax,max(&N0,eRatio**&a*&N0));
         If &DuHa*(uy1-ux1+NId) < 0 Then nFinal = N1;
         If &nAdj = "N" then nFinal = &Nmax;
         If nFinal > N1 Then Do;
             ux2 = Rannor(seedx)*sigma/Sqrt(nFinal-N1)+ux ;
             uy2 = Rannor(seedy)*sigma/Sqrt(nFinal-N1)+uy;
             T2 = (uy2-ux2+NId)*Sqrt(nFinal-N1)/2**0.5/sigma;
             p2=1-ProbNorm(T2);
             If Model="ind" Then TS2=p2;
             If Model="sum" Then TS2=p1+p2;
             If Model="prd" Then TS2=p1*p2;
             If .<TS2<=&alpha2 Then Power=Power+1/&nSims;
         End;
     End;
```

```
    AveN=AveN+nFinal/&nSims;
 End;
 Output;
run;
Proc Print Data=DCSPnormal; run;
%Mend DCSPnormal;
```
‹‹SAS‹‹

Example 4.2 Adaptive Design for Asthma Study

In a phase-III asthma study with 2 dose groups (control and active), the primary efficacy endpoint is the percent change from baseline in FEV1. The estimated FEV1 improvement from baseline is 5% and 12% for the control and active groups, respectively, with a common standard deviation of $\sigma = 22\%$. Based on a large sample assumption, the sample-size for a fixed design is 208 per group, which has 90% power to detect the difference at a one-sided alpha = 0.025. Using MSP, an interim analysis is planned based on the response assessments of 50% of the patients. To design an adaptive trial (GSD), we can use the SAS macro DCSPnormal, described as follows:

(1) Choose stopping boundaries at the first stage: $\alpha_1 = 0.01$, $\beta_1 = 0.15$; then from Table 4.3, we can obtain $\alpha_2 = 0.1871$.

(2) Check the stopping boundary to make sure that the familywise error is controlled by submitting the following SAS statement:

```
%DCSPnormal( Model="sum", alpha=0.025, beta=0.1, sigma=0.22,
ux=0.05, uy=0.05, nInterim=155, Nmax=310, DuHa=0.07, nAdj="N",
alpha1=0.01, beta1=0.15, alpha2=0.1871);
```
‹‹SAS‹‹

The simulated familywise error rate $\alpha = 0.0253$. Therefore the stopping boundaries are confirmed.

(3) Calculate power or sample-size required using the following SAS statement:

```
%DCSPnormal(Model="sum", alpha=0.025, beta=0.1, sigma=0.22,
ux=0.05, uy=0.12, nInterim=155, Nmax=310, DuHa=0.07, nAdj="N",
alpha1=0.01, beta1=0.15, alpha2=0.1871);
```
‹‹SAS‹‹

(4) Perform the sensitivity analysis (under condition H_s) with treatment means of 0.05 and 0.1 for the control and test groups, respectively, by submiting the following SAS statement:

```
>>SAS>>
%DCSPnormal(Model="sum", alpha=0.025, beta=0.1, sigma=0.22,
ux=0.05, uy=0.10, nInterim=155, Nmax=310, DuHa=0.07, nAdj="N",
alpha1=0.01, beta1=0.15, alpha2=0.1871);
<<SAS<<
```

The simulation results are summarized in Table 4.4.

Table 4.4: Operating Characteristics of a GSD with MSP

Simulation condition	FSP	ESP	$\bar{N}$	N_{max}	Power (alpha)
H_o	0.849	0.010	177	310	(0.025)
H_a	0.039	0.682	198	310	0.949
H_s	0.167	0.373	226	310	0.743

From Table 4.4, we can see that the design has a smaller expected sample-size ($\bar{N}$) under H_o and H_a (177, 198) than the classic design (208). If the trial stops early, only 155 patients per group are required. However, the group sequential design has a larger maximum sample-size (310) than the classic design (208). The early futility stopping probability (FSP) and early efficacy stopping probability (ESP) are also shown in Table 4.4. The sensitivity analysis shows a large power loss when treatment difference is slightly lower than expected. To protect power, we can use the sample-size re-estimation method.

Now let's calculate stagewise-ordering adjusted p-values (see Chapter 3). Assume the trial is finished with the stagewise p-value (unadjusted, based on subsample from the stage) for the first stage of $p_1 = 0.012$ (which is larger than $\alpha_1 = 0.01$, therefore the trial continues to the second stage) and the stagewise p-value for the second stage of $p_2 = 0.18$. The test statistic at stage 2 is $t = p_1 + p_2 = 0.012 + 0.18 = 0.192 > \alpha_2 = 0.1871$. Therefore, we failed to reject the null hypothesis and cannot claim superior efficacy of the test drug.

The stagewise-ordering adjusted p-value can be calculated from (4.9):

$$p = \alpha_1 + \frac{1}{2}(t - \alpha_1)^2 = 0.012 + 0.5\,(0.1871 - 0.01)^2 = 0.02\,77.$$

Because $p = 0.02772 > \alpha = 0.025$, we reach the same conclusion: fail to reject the null hypothesis.

4.3 Method with Linear Combination of P-values

For a two-stage design and constant $w_i > 0$, $w_1 = 1$,(for $w_1 \neq 1$, we can rescale w_i such that $w_1 = 1$), we have

$$\pi_1 = \int_0^{\alpha_1} dt = \alpha_1$$

and

$$\pi_2 = \frac{1}{2w_2}\left\{\alpha_1^2 - 2\alpha_1\alpha_2 + 2\alpha_2 \min(\alpha_2, \beta_1) - [\min(\alpha_2, \beta_1)]^2\right\}. \quad (4.10)$$

For $\alpha_2 = \beta_1$, (4.10) becomes $\pi_2 = \frac{1}{2w_2}(\alpha_2 - \alpha_1)^2$ and the α-control requires

$$\alpha_1 + \frac{1}{2w_2}(\alpha_2 - \alpha_1)^2 = \alpha. \quad (4.11)$$

For general line-combination of p-values with a K-stage design, the test statistic is given by

$$T_k = \sum_{i=1}^{k} w_{ki} p_i.$$

The stopping boundary and adjusted p-values can be easily found using computer simulation.

4.4 Method with Product of P-values

This method is referred to as MPP. The test statistic in this method is based on the product of the stagewise p-values from the subsamples. For two-stage designs, the test statistic is defined as

$$T_k = \Pi_{i=1}^{k} p_i, \ k = 1, 2. \quad (4.12)$$

The α spent in the two stages is given by

$$\pi_1 = \int_0^{\alpha_1} dt_1 = \alpha_1 \quad (4.13)$$

and

$$\pi_2 = \int_{\alpha_1}^{\beta_1} \int_0^{\alpha_2} \frac{1}{t_1} dt_2 dt_1. \quad (4.14)$$

Carrying out the integrations in (4.14) and substituting the results into (3.7), we can obtain the following formulation for determining stopping boundaries:

$$\alpha = \alpha_1 + \alpha_2 \ln \frac{\beta_1}{\alpha_1},\ \alpha_1 < \beta_1 \leq 1. \tag{4.15}$$

Note that the stopping boundaries based on Fisher's criterion are special cases of (4.15), where $\alpha_2 = \exp\left[-\frac{1}{2}\chi_4^2(1-\alpha)\right]$, i.e., $\alpha_2 = 0.00380$ for $\alpha = 0.025$. To calculate the stopping boundaries, one can predetermine α, α_1, and β_1, then solve (4.15) for α_2. See Table 4.5 for examples.

Table 4.5: Stopping Boundaries α_2 with MPP

	α_1	0.001	0.0025	0.005	0.010	0.015	0.020
β_1							
0.15		0.0048	0.0055	0.0059	0.0055	0.0043	0.0025
0.20		0.0045	0.0051	0.0054	0.0050	0.0039	0.0022
0.25		0.0043	0.0049	0.0051	0.0047	0.0036	0.0020
0.30	α_2	0.0042	0.0047	0.0049	0.0044	0.0033	0.0018
0.35		0.0041	0.0046	0.0047	0.0042	0.0032	0.0017
0.40		0.0040	0.0044	0.0046	0.0041	0.0030	0.0017
0.50		0.0039	0.0042	0.0043	0.0038	0.0029	0.0016
1.00		0.0035	0.0038	0.0038	0.0033	0.0024	0.0013

Note: One-sided $\alpha = 0.025$.

The stagewise-ordering p-value can be obtained using

$$p(t;k) = \begin{cases} t, & k = 1, \\ \alpha_1 + t \ln \frac{\beta_1}{\alpha_1}, & k = 2\ , \end{cases} \tag{4.16}$$

where $t = p_1$ if the trial stops at stage 1 ($k = 1$) and $t = p_1 p_2$ if the trial stops at stage 2 ($k = 2$).

It is interesting to know that when $p_1 < \alpha_2$, there is no point in continuing the trial because $p_1 p_2 < p_1 < \alpha_2$ and efficacy should be claimed. Therefore it is suggested that we should choose $\beta_1 > \alpha_2$ and $\alpha_1 > \alpha_2$.

The SAS Macro 4.3 has been implemented for simulating two-arm adaptive designs with survival endpoint. The adaptive method can be based on individual stagewise p-values, or the sum or product of the stagewise p-values. The SAS variables are defined as follows: **ux** and **uy** are true hazard rates in the x and y groups, respectively. **DuHa** = the estimate for the true treatment difference under the alternative H_a, and **N** = sample-size per group. **Alpha1** = early efficacy stopping boundary (one-sided),

beta1 = early futility stopping boundary (one-sided), and **alpha2** = final efficacy stopping boundary (one-sided). The null hypothesis test is H_o: $\delta +$ **NId** < 0, where δ = **uy** - **ux** is the treatment difference and **NId** = noninferiority margin. **nSims** = the number of simulation runs, **alpha** = one-sided overall type-I error rate, and **beta** = type-II error rate. **nAdj** = "N" for the case without sample-size re-estimation and **nAdj** = "Y" for the case with sample-size adjustment, **Nmax** = maximum sample-size allowed; **N0** = the initial sample-size at the final analysis; **nInterim** = sample-size for the interim analysis; **a** = the parameter in (4.17) for the sample-size adjustment; **FSP** = futility stopping probability; **ESP** = efficacy stopping probability; **AveN** = average sample-size; **Power** = power of the hypothesis test; **nClassic** = sample-size for the corresponding classic design; **Model** = "ind", "sum", or "prd" for the methods, MIP, MSP, and MPP, respectively.

>>SAS Macro 4.3: Two-Stage Adaptive Design with Survival Endpoint>>

```
%Macro DCSPSurv(nSims=1000000, model="sum", alpha=0.025,
        beta=0.2, NId=0, tStd=12, tAcr=4, ux=0, uy=1, DuHa=1,
        nAdj="Y", Nmax=100, N0=100, nInterim=50, a=2,
        alpha1=0.01, beta1=0.15, alpha2=0.1871);
Data DCSPSurv; Keep Model FSP ESP AveN Power nClassic;
 seedx=2534; seedy=6762; alpha=&alpha; NId=&NId;
 Nmax=&Nmax; ux=&ux; uy=&uy; N1=&nInterim; model=&model;
 Expuxd=exp(-ux*&tStd);  Expuyd=exp(-uy*&tStd);
 sigmax=ux*(1+Expuxd*(1-exp(ux*&tAcr))/(&tAcr*ux))**(-0.5);
 sigmay=uy*(1+Expuyd*(1-exp(uy*&tAcr))/(&tAcr*uy))**(-0.5);
 sigma=((sigmax**2+sigmay**2)/2)**0.5;
 eSize=abs(&DuHa+NId)/sigma;
 nClassic=Round(2*((probit(1-alpha)+Probit(1-&beta))/eSize)**2);
 FSP=0; ESP=0; AveN=0; Power=0;
 Do isim=1 To &nSims;
 nFinal=N1;
    ux1 = Rannor(seedx)*sigma/Sqrt(N1)+ux;
    uy1 = Rannor(seedy)*sigma/Sqrt(N1)+uy;
    T1 = (uy1-ux1+NId)*Sqrt(N1)/2**0.5/sigma;
    p1=1-ProbNorm(T1);
    If p1>&beta1 Then FSP=FSP+1/&nSims;
    If p1<=&alpha1 Then do;
        Power=Power+1/&nSims; ESP=ESP+1/&nSims;
    End;
```

```
    If p1>&alpha1 and p1<=&beta1 Then Do;
        eRatio=Abs(&DuHa/(Abs(uy1-ux1)+0.0000001));
        nFinal=min(Nmax,max(&N0,eRatio**&a*&N0));
        If &DuHa*(uy1-ux1+NId)<0 then nFinal=N1;
        If &nAdj="N" then nFinal=Nmax;
        If nFinal>N1 Then Do;
            ux2 = Rannor(seedx)*sigma/Sqrt(nFinal-N1)+ux ;
            uy2 = Rannor(seedy)*sigma/Sqrt(nFinal-N1)+uy;
            T2 = (uy2-ux2+NId)*Sqrt(nFinal-N1)/2**0.5/sigma;
            p2=1-ProbNorm(T2);
            If Model="ind" Then TS2=p2;
            If Model="sum" Then TS2=p1+p2;
            If Model="prd" Then TS2=p1*p2;
            If .<TS2<=&alpha2 Then Power=Power+1/&nSims;
        End;
    End;
    AveN=AveN+nFinal/&nSims;
 End;
 Output;
Run;
Proc Print Data=DCSPSurv; Run;
%Mend DCSPSurv;
```

‹‹SAS‹‹

Example 4.3 Adaptive Design for Oncology Trial

In a two-arm comparative oncology trial, the primary efficacy endpoint is time-to-progression (TTP). The median TTP is estimated to be 8 months (hazard rate = 0.08664) for the control group, and 10.5 months (hazard rate = 0.06601) for the test group. Assume a uniform enrollment with an accrual period of 9 months and a total study duration of 24 months. The log-rank test will be used for the analysis. An exponential survival distribution is assumed for the purpose of sample-size calculation. The classic design requires a sample-size of 323 subjects per group.

We design the trial with one interim analysis when 40% of patients have been enrolled. The interim analysis for efficacy is planned based on TTP, but it does not allow for futility stopping. Using MPP, we choose the following boundaries: $\alpha_1 = 0.005$, $\beta_1 = 1$ ($\beta_1 = 1$ implies no futility stopping), and $\alpha_2 = 0.0038$ from Table 4.5. Again, we follow the same steps as for the two previous examples using the SAS macro **DCSPSurv**: Note that in SAS Macro 4.3, we again assume that the stagewise p-values are mutually independent. The steps for the simulations are:

(1) Choose stopping boundaries at the first stage: $\alpha_1 = 0.005$, $\beta_1 = 1$; then from Table 4.5, we can obtain $\alpha_2 = 0.0038$.

(2) Check the stopping boundary to make sure that the familywise error is controlled by submitting the following SAS statement:

```
>>SAS>>
%DCSPSurv(model="prd", alpha=0.025, beta=0.15, tStd=24, tAcr=9,
ux=0.08664, uy=0.08664, DuHa=0.02063,  nAdj="N", Nmax=344,
N0=344, nInterim=138, alpha1=0.005, beta1=1, alpha2=0.0038);
<<SAS<<
```

The simulated familywise error rate $\alpha = 0.0252$. Therefore the stopping boundaries are confirmed.

(3) Calculate power or sample-size required using the following SAS macro calling:

```
>>SAS>>
%DCSPSurv(model="prd", alpha=0.025, beta=0.15, tStd=24, tAcr=9,
ux=0.06601, uy=0.08664, DuHa=0.02063, nAdj="N", Nmax=344,
N0=344, nInterim=138, alpha1=0.005, beta1=1, alpha2=0.0038);
<<SAS<<
```

We modified the sample-size until it reached the desired power. It turns out that the maximum sample-size is 344 and the sample-size for the interim analysis is 138 per group.

(4) Perform a sensitivity analysis under the condition H_s. Because a 2.5-month difference in median TTP is a conservative estimate, the obvious question is what is the early stopping probability if the true treatment difference in median TTP is, for example, 3 months (8 months versus 11 months or hazard rate = 0.06301)? To answer this question, we issue the following SAS statement with hazard rates of 0.08664 and 0.06301 for the control and test groups, respectively.

```
>>SAS>>
%DCSPSurv(model="prd", alpha=0.025, beta=0.15, tStd=24, tAcr=9,
ux=0.06301, uy=0.08664, DuHa=0.02363, nAdj="N",  Nmax=344,
N0=344, nInterim=138, alpha1=0.005, beta1=1, alpha2=0.0038);
<<SAS<<
```

We now summarize the simulation outputs for the three scenarios in Table 4.6.

Table 4.6: Operating Characteristics of a GSD with MPP

Simulation condition	FSP	ESP	$\bar{N}$	N_{max}	Power (alpha)
H_o	0	0.005	343	344	0.025
H_a	0	0.268	289	344	0.851
H_s	0	0.381	265	344	0.937

From Table 4.6, we can see that the design has a smaller expected sample-size ($\bar{N}$) under H_a (289/group) than the classic design (323/group). If the trial stops early, only 238 patients per group are required. However, the group sequential design has a larger maximum sample-size (344) than the classic design (323). The early futility stopping probability (FSP) and early efficacy stopping probability (ESP) are also shown in Table 4.6. The sensitivity analysis shows that the early stopping probability increases from 26.8% to 38.1% if the difference in median TTP is 3 months instead of 2.5 months. This indicates a time savings, too. Also, the power will be 93.7% if the difference in median TTP is 3 months instead of 2.5 months. We can see that the group sequential design is very advantageous when the effect size is larger than our initial estimate. We will discuss this in a later chapter on choosing adaptive designs.

Now let's calculate adjusted p-values. If the trial is stopped at the first stage with $p_1 = 0.002$, then the conditional and overall p-values are the same and equal to 0.002. If the first stagewise p-value $p_1 = 0.05$ (which is larger than $\alpha_1 = 0.002$ and not significant; therefore the trial continued to the second stage) and $p_2 = 0.07$, the test statistic at stage 2 is $t = p_1 p_2 = (0.05)(0.07) = 0.0035 < \alpha_2 = 0.0038$. Therefore, we reject the null hypothesis and claim the efficacy of the test drug. The stagewise-ordering p-value can be calculated from (4.16):

$$p = \alpha_1 + t \ln \frac{\beta_1}{\alpha_1} = 0.005 + 0.0185 = 0.0235.$$

Example 4.4 Early Futility Stopping Design with Binary Endpoint

We use an early example. A phase III trial is to be designed for patients with acute ischemic stroke of recent onset. The composite endpoint (death and MI) is the primary endpoint and event rate is 14% for the control group and 12% for the test group. Based on a large sample assumption, the sample-size for a fixed design is 5937 per group, which provides 90% power to detect the difference at one-sided alpha = 0.025. An interim analysis for futility stopping is planed based on 50% patients' response assessments. The interim look is for futility stopping. We can use both MIP

and MSP, but we don't recommend MPP in this case because MPP doesn't allow for futility early stopping only. Because the regulatory agency may be concerned that in the current practice, the futility boundary may not be followed, i.e., the trial did continue when in fact the futility boundary is crossed, to protect type-I error, it was suggested that the futility boundary should not be used for determining the stopping boundaries at later stages. We know that MIP and MSP have used the futility boundary in determination of the subsequent boundaries. However, we will propose a better procedure in which the stopping boundaries are different depending on whether the futility boundary is followed or not. Let's illustrate this with MSP for this trial.

(1) Choose futility stopping boundaries from Table 4.5: $\alpha_1 = 0$ (It implies no early efficacy stopping), $\beta_1 = 0.15,\ \alpha_2 = 0.2417$. If the futility boundary has not been followed, the first stage subsample will not be used in the final analysis and $\alpha_2 = 0.025$ will be used. Alternatively, we can conservatively use $\alpha_1 = 0,\ \beta_1 = 0.15,\ \alpha_2 = 0.2236$ (This α_2 is corresponding to $\alpha_1 = 0$ and $\beta_1 = 1$).

(2) Run simulations to obtain the sample-size required and the operating characteristics under H_a with $\alpha = 0.025$ and $\beta = 0.1$ (power = 0.9). By trying different maximum sample-size (N_{max}) in the following SAS statement, we found that 7360 gives the desired power (the classic design requires n = 5937 per group):

```
>>SAS>>
 %DCSPbinary(Model="sum", alpha=0.025, beta=0.1, Px=0.12, Py=0.14,
 DuHa=0.02, Nmax=7360, nInterim=3680, alpha1=0, beta1=0.15,
alpha2=0.2417);
<<SAS<<
```

(3) To obtain the operating characteristics under H_o, submit the following SAS statement:

```
>>SAS>>
%DCSPbinary(Model="sum", alpha=0.025, beta=0.1, Px=0.14, Py=0.14,
DuHa=0.02, Nmax=7360, nInterim=3680, alpha1=0, beta1=0.15,
alpha2=0.2417);
<<SAS<<
```

The simulation results are presented in Table 4.7.

Table 4.7: Operating Characteristics of a GSD with MPP

Simulation condition	FSP	ESP	$\bar{N}$	N_{max}	Power (alpha)
H_o	0.851	0	4232	7360	0.025
H_a	0.065	0	7124	7360	0.899

Note that the futility design is used because there is a great concern that the drug may not have efficacy. In such a case, the expected sample-size is 4229/group which is much smaller than the classic design (5937/group).

If the conservative stopping boundaries ($\alpha_1 = 0$, $\beta_1 = 0.15$, and $\alpha_2 = 0.2236$) are used, the simulated power is 89.4% with the same sample-size (Nmax = 7360) by submitting the following SAS statement:

```
>>SAS>>
%DCSPbinary(Model="sum", alpha=0.025, beta=0.1, Px=0.12, Py=0.14,
DuHa=0.02, Nmax=7360,  nInterim=3680, alpha1=0, beta1=0.15,
alpha2=0.2236);
<<SAS<<
```

Using the conservative boundaries, the α is actually controlled at a 0.0224 level. In this example, there is minimal difference in power between the two methods.

Example 4.5 Noninferiority Design with Binary Endpoint

Let's consider a noninferiority/superiority design for the trial in Example 4.4. If superiority is not achieved, we will perform a noninferiority test. Because of the closed testing procedure, no alpha adjustment is required for the two hypothesis tests. The noninferiority boundary is decided to be 0.5%. For the purpose of comparison, we use the same sample-size and stopping boundaries as in Example 4.4.

(1) Choose futility stopping boundaries from Table 4.5: $\alpha_1 = 0$, $\beta_1 = 0.15$, and $\alpha_2 = 0.2417$.

(2) Perform simulations by using the following SAS statement to obtain the sample-size required and the operating characteristics under H_a with $\alpha = 0.025$ and Nmax = 7360 per group:

```
>>SAS>>
%DCSPbinary(Model="sum", alpha=0.025, beta=0.1,  NId=0.005,
Px=0.12, Py=0.14, DuHa=0.02, Nmax=7360,  nInterim=3680, alpha1=0,
beta1=0.15, alpha2=0.2417);
<<SAS<<
```

(3) Obtain the operating characteristics under Ho by running simulations under H_o:

```
>>SAS>>
%DCSPbinary(Model="sum", alpha=0.025, beta=0.1,  NId=0.005,
Px=0.145, Py=0.14, DuHa=0.02, Nmax=7360,  nInterim=3680, a=2,
alpha1=0, beta1=0.15, alpha2=0.2417);
<<SAS<<
```

Note that the futility design is used because there is great concern that the drug may not have efficacy. In such a case, the expected sample-size is 4923/group.

(4) Perform the sensitivity analysis under the condition that the event rate is 12.5%. The power for the noninferiority test under this condition is 89.5%, which is obtained by submitting the following SAS statement.

```
>>SAS>>
%DCSPbinary(Model="sum", alpha=0.025, beta=0.1, NId=0.005,
Px=0.125, Py=0.14, DuHa=0.02, Nmax=7360, nInterim=3680, alpha1=0,
beta1=0.15, alpha2=0.2417);
<<SAS<<
```

The simulation results are presented in Table 4.8.

Table 4.8: Operating Characteristics of a GSD with MPP

Simulation condition	FSP	ESP	$\bar{N}$	N_{max}	Power (alpha)
H_o	0.850	0	4232	7360	0.025
H_a	0.010	0	7302	7360	0.977
H_s	0.068	0	7110	7360	0.895

Example 4.6 Sample-Size Re-estimation with Normal Endpoint

In a phase-III asthma study with 2 dose groups (control and active), the primary efficacy endpoint is the percent change from baseline in FEV1. The estimated FEV1 improvement from baseline is 5% and 12% for the control and active groups, respectively, with a common standard deviation of $\sigma = 22\%$. Based on a large sample assumption, the sample-size for a fixed design is 208 per group with 95% power and a one-sided alpha = 0.025. Using MSP, an interim analysis is planned based on the response assessments of 50% of the patients. The interim analysis is used for sample-size re-estimation and also for futility stopping. We follow the steps below to design the adaptive trial.

(1) Choose stopping boundaries at the first stage: $\alpha_1 = 0,\ \beta_1 = 0.25$, then from Table 4.3, we obtain $\alpha_2 = 0.2236$.

(2) Determine the rule for sample-size re-estimation:

$$N = \left(\frac{E_0}{E}\right)^a N_0, \tag{4.17}$$

where N is the newly estimated sample-size, N_0 = initial sample-size, a is a constant and often chosen to be 2, and E_0 and E are predetermined and observed effect sizes or treatment differences, respectively. Choose $N_0 = 242$, which is suggested to be close to but larger than the sample-size for the classic design. The choice of N_0 should be dependent on the operating characteristics of the design, therefore, several iterations may be required before a satisfactory N_0 is chosen.

(3) Perform simulations without SSR using the following SAS statement:

```
>>SAS>>
%DCSPnormal(Model="sum", alpha=0.025, beta=0.1, sigma=0.22,
ux=0.05, uy=0.12, nInterim=121, Nmax=242, N0=242, DuHa=0.07,
alpha1=0, beta1=0.25, alpha2=0.2236);
<<SAS<<
```

(4) Perform simulations with SSR using the following SAS statement:

```
>>SAS>>
 %DCSPnormal(Model="sum", alpha=0.025, beta=0.1, sigma=0.22,
ux=0.05, uy=0.12, nInterim=121, Nmax=350, N0=242, DuHa=0.07,
nAdj="Y", alpha1=0, beta1=0.25, alpha2=0.2236);
<<SAS<<
```

(5) Perform simulations for sensitivity analysis without SSR using the following SAS statement:

```
>>SAS>>
 %DCSPnormal(Model="sum", alpha=0.025, beta=0.1, sigma=0.22,
ux=0.05, uy=0.105, nInterim=121, Nmax=242, N0=242, DuHa=0.07,
alpha1=0, beta1=0.25, alpha2=0.2236);
<<SAS<<
```

(6) Perform simulations for sensitivity analysis with SSR (note that DuHa = 0.07 for sensitivity analysis):

>>SAS>>
%DCSPnormal(Model="sum", alpha=0.025, beta=0.1, sigma=0.22, ux=0.05, uy=0.105, nInterim=121, Nmax=350, N0=242, DuHa=0.07, nAdj="Y", alpha1=0, beta1=0.25, alpha2=0.2236);
<<SAS<<

The simulation results are presented in Table 4.9. Note that we set the futility boundary at $\beta_1 = 0.25$ due to the consideration that if the treatment effect is very small (such that $p_1 > 0.25$), the required sample-size to be adjusted is very large and not feasible.

Table 4.9: Operating Characteristics of a GSD with MSP

Simulation condition	Method	FSP	$\bar{N}$	N_{max}	Power
H_o	classic	0	208	208	0.025
	without SSR	0.750	151	242	0.025
	with SSR	0.750	177	350	0.025
H_a	classic	0	208	208	0.900
	without SSR	0.036	238	242	0.900
	with SSR	0.036	278	350	0.928
H_s	classic	0	208	208	0.722
	without SSR	0.102	230	242	0.733
	with SSR	0.102	285	350	0.804

From Table 4.9, we can see that the adaptive design has a smaller expected sample-size ($\bar{N}$) under H_o than the classic design (208). When using the n-re-estimation mechanism, the power is protected to a certain degree (72.2% for the classic vs. 73.3% for adaptive design without SSR and 80.4% with SSR). Of course, this power protection is at the cost of sample-size. Note the average n = 285/group with sample-size adjustment when the effect is 5% versus 10.5%. If this sample-size is used in a classic design, the power would be 84.7%. The sample-size re-estimation has lost its efficiency in this sense, though there is a saving in sample-size under H_o.

Example 4.7 Sample-Size Re-estimation with Survival Endpoint

In this example we will compare MIP, MSP, and MPP and illustrate how to calculate the adjusted p-values with these 3 different methods.

Suppose in a two-arm comparative oncology trial, the primary efficacy endpoint is time to progression (TTP). The median TTP is estimated to be 8 months (hazard rate = 0.08664) for the control group, and 10.5 months (hazard rate = 0.06601) for the test group. Assume uniform enrollment

with an accrual period of 9 months and a total study duration of 24 months. The log-rank test will be used for the analysis. An exponential survival distribution is assumed for the purpose of sample-size calculation.

To generate the operating characteristics using MIP, MSP, and MPP, we use the following SAS macro calls, respectively:

```
>>SAS>>
%DCSPsurv(model="ind", alpha=0.025, beta=0.15,  tStd=24, tAcr=9,
 ux=0.06601, uy=0.08664, DuHa=0.02063, nAdj="Y",  Nmax=400,
N0=350, nInterim=200, alpha1=0.01, beta1=0.25, alpha2=0.0625);

%DCSPsurv(model="sum", alpha=0.025, beta=0.15, NId=0, tStd=24,
tAcr=9, ux=0.06601, uy=0.08664, DuHa=0.02063, nAdj="Y",  Nmax=400,
 N0=350, nInterim=200, alpha1=0.01, beta1=0.25, alpha2=0.1832);

%DCSPsurv(model="prd", alpha=0.025, beta=0.15, tStd=24, tAcr=9,
ux=0.06601, uy=0.08664, DuHa=0.02063, nAdj="Y", Nmax=400, N0=350,
 nInterim=200, a=2, alpha1=0.01, beta1=0.25, alpha2=0.00466);
<<SAS<<
```

When there is a 10.5-month median time for the test group, the classic design requires a sample-size of 323 per group with 85% power at a level of significance (one-sided) $\alpha = 0.025$. To increase efficiency, an adaptive design with an interim sample-size of 200 patients per group is used. The interim analysis allows for early efficacy or futility stopping with stopping boundaries (from Tables 4.1, 4.3, and 4.5) $\alpha_1 = 0.01$, $\beta_1 = 0.25$, and $\alpha_2 = 0.0625$ for MIP, 0.1832 for MSP, and 0.00466 for MPP. The sample-size adjustment is based on (4.17). The maximum sample-size allowed for adjustment is $N_{\max} = 400$. The parameter for sample-size adjustment N_o is 350 (N_o is usually chosen to be close to the sample-size from the classic design so that the adaptive design will have similar power to the classic design.). The simulation results are presented in Table 4.10, where the abbreviations ESP and FSP stand for early efficacy stopping probability and early futility stopping probability, respectively.

Table 4.10: Operating Characteristics of Adaptive Methods

Median time				Expected	Power (%)
Test	Control	ESP	FSP	N	MIP/MSP/MPP
8	8	0.010	0.750	248	2.5/2.5/2.5
10.5	8	0.512	0.046	288	86.3/87.3/88.8

Note: 1,000,000 simulation runs.

Note that power is the probability of rejecting the null hypothesis. Therefore when the null hypothesis is true, the power is the type-I error rate α. From Table 4.10, it can be seen that the one-sided α is controlled at a 0.025 level as expected for all three methods. The expected sample sizes under both H_o and H_a are smaller than the sample-size for the classic design (290/group). The power is 86.3%, 87.3%, and 88.8% for MIP, MSP, and MPP, respectively. All three designs have the same expected sample-size of 288/group which is smaller than the sample-size (323/group) for the classic design with 85% power. In adaptive design, conditional power is more important than power. We will discuss this in detail later.

4.5 Event-Based Adaptive Design

The methods discussed for survival analyses so far are based on the number of patients at each stage, instead of number of events. The reason for this is that the methods are based on the assumption of independent stagewise statistics. Therefore, the first N_1 patients enrolled will be used for the first interim analysis regardless they have the event or not. Strictly speaking, the commonly used log-rank test statistics based on number of events,

$$T\left(\hat{D}_k\right) = \sqrt{\frac{\hat{D}_k}{2}} \ln \frac{\hat{\lambda}_1}{\hat{\lambda}_2} \sim N\left(\sqrt{\frac{D_k}{2}} \ln \frac{\lambda_1}{\lambda_2}, 1\right), \tag{4.18}$$

are not independent, where D_k is the number of events at the k^{th} stage. However, Breslow and Haug (1977) and Canner (1997) showed that the independent normal approximation works well for small D_k. The relationship between the number of deaths and number of patients is given in Problem 2.3 in Chapter 2. Using (4.18), we can implement adaptive design for survival based on the number of events as follows.

The SAS Macro 4.4 has been implemented for simulating two-arm group sequential designs with survival endpoint. The adaptive method can be based on individual stagewise p-values, or the sum or product of the stagewise p-values. The SAS variables are defined as follows: **ux** and **uy** are true hazard rates in the x and y groups, respectively. **N** = sample-size per group. **Alpha1** = early efficacy stopping boundary (one-sided), **beta1** = early futility stopping boundary (one-sided), and **alpha2** = final efficacy stopping boundary (one-sided). **nSims** = the number of simulation runs, **alpha** = one-sided overall type-I error rate, and **beta** = type-II error rate. **InfoTime** = sample-size ratio for the interim analysis; **FSP** = futility stopping probability; **ESP** = efficacy stopping probability; **AveDs**

= average total number of events; **Power** = power of the hypothesis test; **Model** = "ind", "sum", or "prd" for the methods, MIP, MSP, and MPP, respectively.

››SAS Macro 4.4: Event-Based Adaptive Design››

```
%Macro DCSPSurv2(nSims=1000000, model="sum", alpha=0.025,
        beta=0.2, tStd=12, tAcr=4, ux=0.08, uy=0.1, N=100,
        InfoTime=0.5, alpha1=0.01, beta1=0.15, alpha2=0.1871);
Data DCSPSurv; Keep Model FSP ESP Power AveDs N;
seed1=2534; seed2=2534; alpha=&alpha; N=&N; ux=&ux;
    uy=&uy; model=&model; infoTime=&infoTime; tAcr=&tAcr;
FSP=0; ESP=0; AveDs=0; Power=0; u=(ux+uy)/2;
    Ds=2*N/&tAcr*(&tAcr-(exp(u*&tAcr)-1)/u*exp(-u*&tStd));
    Ds1=Ds*infoTime;
    nFinal=Ds1;
Do isim=1 To &nSims;
     T1 = Rannor(seed1)+Sqrt(Ds1/2)*log(uy/ux);
     p1=1-ProbNorm(T1);
     If p1>&beta1 Then FSP=FSP+1/&nSims;
     If p1<=&alpha1 Then do;
          Power=Power+1/&nSims; ESP=ESP+1/&nSims;
     End;
     If p1>&alpha1 and p1<=&beta1 Then Do;
          nFinal=Ds;
          T2 = Rannor(seed2)+Sqrt((Ds-Ds1)/2)*log(uy/ux);
          p2=1-ProbNorm(T2);
          If Model="ind" Then TS2=p2;
          If Model="sum" Then TS2=p1+p2;
          If Model="prd" Then TS2=p1*p2;
          If .<TS2<=&alpha2 Then Power=Power+1/&nSims;
     End;
     AveDs=AveDs+nFinal/&nSims;
End;
Output;
Run;
Proc Print Data=DCSPSurv; Run;
%Mend DCSPSurv2;
```

‹‹SAS‹‹

An example of using this SAS macro is presented as follows:

››SAS››

```
%DCSPSurv2(nSims=100000, model="sum", alpha=0.025, beta=0.2,
        tStd=24, tAcr=9,  ux=0.06601, uy=0.08664, N=180,
        InfoTime=0.5, alpha1=0.01, beta1=0.15, alpha2=0.1871);
```
‹‹**SAS**‹‹

Whether based on the number of events or patients, the results are very similar. SAS Macro 4.4 can be extended to general adaptive design with sample-size re-estimation (Problem 4.3). Other methods for adaptive design with a survival endpoint can be found from work by Li, Shih, and Wang (2005). Practically, the accrual has to continue in most cases when collecting the data and performing the interim analysis; it is often the case that at the time when interim analysis is done, most or all patients are enrolled. What is the point to have the interim analysis? The answer is a positive interim analysis would allow the drug to be on the market earlier.

4.6 Adaptive Design for Equivalence Trial

In Chapter 2, we have studied the equivalence test for the two parallel groups:

$$H_0 : |\mu_T - \mu_R| \geq \delta \quad \text{versus} \quad H_a : |\mu_T - \mu_R| < \delta. \tag{4.19}$$

If the null hypothesis is rejected, then we conclude that the test drug and the reference drug are equivalent.

For a large sample-size, the null hypothesis is rejected if

$$T_1 = \frac{\bar{x}_R - \bar{x}_T - \delta}{\hat{\sigma}\sqrt{\frac{2}{n}}} < -z_{1-\alpha} \text{ and } T_2 = \frac{\bar{x}_R - \bar{x}_T + \delta}{\hat{\sigma}\sqrt{\frac{2}{n}}} > z_{1-\alpha}. \tag{4.20}$$

The approximate sample-size per group is given by (see Chapter 2)

$$n = \frac{2\left(z_{1-\alpha} + z_{1-\beta/2}\right)^2 \sigma^2}{\left(|\varepsilon| - \delta\right)^2}, \tag{4.21}$$

(4.20) is equivalent to the following condition:

$$p = \max\left\{p_{01}, p_{02}\right\} \leq \alpha \tag{4.22}$$

where $p_{01} = \Phi\left(T_1\right)$ and $p_{02} = \Phi\left(-T_2\right)$.

We now discuss the two-stage adaptive design that allows for sample-size adjustment based on information at the first stage. The key for the adaptive equivalence trial is to define an appropriate stagewise p-value, Chang (2007) suggests using the p-value similar to (4.22) but based on subsample from the k^{th} stage, i.e.,

$$p_k = \max\left\{\Phi\left(T_{k1}\right), \Phi\left(-T_{k2}\right)\right\}, \tag{4.23}$$

where

$$T_{k1} = \frac{\bar{x}_{kR} - \bar{x}_{kT} - \delta}{\hat{\sigma}_k\sqrt{\frac{2}{n_k}}} \text{ and } T_{k2} = \frac{\bar{x}_{kR} - \bar{x}_{kT} + \delta}{\hat{\sigma}_k\sqrt{\frac{2}{n_k}}}. \tag{4.24}$$

Using the stagewise p-values defined in (4.23), we can use MIP, MSP, and MPP to design adaptive equivalence trial without any difficulty. For convenience, let's implement the method in SAS.

The SAS Macro 4.5 has been implemented for simulating two-arm equivalence trial with normal endpoint. The adaptive method can be based on individual stagewise p-values, or the sum or product of the stagewise p-values. The SAS variables are defined as follows: **ux** and **uy** are true proportions of response in the x and y groups, respectively. **DuHa** = the estimate for the true treatment difference under the alternative H_a, and **N** = sample-size per group. **alpha1** = early efficacy stopping boundary (one-sided), **beta1** = early futility stopping boundary (one-sided), and **alpha2** = final efficacy stopping boundary (one-sided). The null hypothesis test is given by (4.19). **NId** = equivalence margin, **nSims** = the number of simulation runs, **alpha** = one-sided overall type-I error rate, and **beta** = type-II error rate. **nAdj** = "N" for the case without sample-size re-estimation and **nAdj** = "Y" for the case with sample-size adjustment, **Nmax** = maximum sample-size allowed; **N0** = the initial sample-size at the final analysis; **nInterim** = sample-size for the interim analysis; **a** = the parameter in (4.17) for the sample-size adjustment; **FSP** = futility stopping probability; **ESP** = efficacy stopping probability; **AveN** = average sample-size; **Power** = power of the hypothesis testing; and **Model** = "ind", "sum", or "prd" for the methods MIP, MSP, and MPP, respectively.

>>SAS Macro 4.5: Adaptive Equivalence Trial Design>>

```
%Macro DCSPEqNormal(nSims=1000000, Model="sum", alpha=0.05,
beta=0.2, sigma=0.3, NId=0.2, ux=0, uy=0.1, nInterim=50, Nmax=100,
N0=100, DuHa=1, nAdj="Y", a= -2, alpha1=0, beta1=0.2, alpha2=0.3);
```

```
Data DCSPEqNormal; Keep Model FSP ESP AveN Power;
seedx=1736; seedy=6214; alpha=&alpha; NId=&NId; Nmax=&Nmax;
ux=&ux; uy=&uy; sigma=&sigma; model=&Model; N1=&nInterim;
eSize=abs(&DuHa+NId)/sigma;
FSP=0; ESP=0; AveN=0; Power=0;
Do isim=1 To &nSims;
    nFinal=N1;
    ux1 = Rannor(seedx)*sigma/Sqrt(N1)+ux;
    uy1 = Rannor(seedy)*sigma/Sqrt(N1)+uy;
    T11 = (uy1-ux1-NId)*sqrt(N1/2)/sigma;
    T12 = (uy1-ux1+NId)*sqrt(N1/2)/sigma;
    p11=Probnorm(T11);
    p12=Probnorm(-T12);
    p1=max(p11,p12);

    If p1>&beta1 then FSP=FSP+1/&nSims;
    If p1<=&alpha1 then do;
        Power=Power+1/&nSims; ESP=ESP+1/&nSims;
    End;
    If p1>&alpha1 and p1<=&beta1 Then Do;
        eRatio = abs(&DuHa/(abs(uy1-ux1)+0.0000001));
        nFinal = min(&Nmax,max(&N0,eRatio**&a*&N0));
        If &DuHa*(uy1-ux1+NId) < 0 Then nFinal = N1;
        If &nAdj = "N" then nFinal = &Nmax;
        If nFinal > N1 Then Do;
            ux2 = Rannor(seedx)*sigma/sqrt(nFinal-N1)+ux ;
            uy2 = Rannor(seedy)*sigma/sqrt(nFinal-N1)+uy;
            T21 = (uy2-ux2-NId)*sqrt(nFinal-N1)/2**0.5/sigma;
            T22 = (uy2-ux2+NId)*sqrt(nFinal-N1)/2**0.5/sigma;
            p21=Probnorm(T21);
            p22=Probnorm(-T22);
            p2=max(p21,p22);
            If Model="ind" Then TS2=p2;
            If Model="sum" Then TS2=p1+p2;
            If Model="prd" Then TS2=p1*p2;
            If .<TS2<=&alpha2 Then Power=Power+1/&nSims;
        End;
    End;
    AveN=AveN+nFinal/&nSims;
End;
Output;
```

```
run;
Proc Print Data=DCSPEqNormal; run;
%Mend DCSPEqNormal;
‹‹SAS‹‹
```

This SAS Macro can also be used for binary and survival endpoints as long as one provides the corresponding "standard deviation" as shown in Table 2.1.

Example 4.8 Adaptive Equivalence LDL Trial

We use the LDL trial in Example 2.2; the equivalence margin is assumed to be $\delta = 5\%$; the treatment difference in LDL is 1% (70% versus 71%) with a standard deviation of 30%. Suppose we decide to use $\alpha = 0.05$ and initial sample-size $N_0 = 1200$ per group, the maximum sample-size $N_{max} = 2000$ per group, and the interim analysis sample-size $N_1 = 600$ per group. The SSR algorithm is given by (4.17) with the parameter $a = -2$. Using MSP, we choose $\alpha_1 = 0$ and $\beta_1 = 0.2$, then $\alpha_2 = 0.35$ (from Eq.(4.8)).

To study the operating characteristics, we use the following SAS macro calls for the design:

```
››SAS››
Title " Check alpha under Ho with alpha = 0.05";
%DCSPEqNormal(Model="sum", alpha=0.05, beta=0.2, sigma=.3,
NId=0.05, ux=0.70, uy=0.75, nInterim=600, Nmax=2000, N0=1200,
DuHa=0.01, nAdj="Y", a=-2, alpha1=0.0, beta1=.2, alpha2=0.35);

Title " Simulate the Power under Ha";
%DCSPEqNormal(Model="sum", alpha=0.05, beta=0.2, sigma=.3,
NId=0.05, ux=0.70, uy=0.71, nInterim=600, Nmax=2000, N0=1200,
DuHa=0.01, nAdj="Y", a=-2, alpha1=0.0, beta1=.2, alpha2=0.35);

Title " Sensitivity Analysis (without SSR)";
%DCSPEqNormal(Model="sum", alpha=0.05, beta=0.2, sigma=.3,
NId=0.05, ux=0.70, uy=0.72, nInterim=600, Nmax=1200, N0=1200,
DuHa=0.01, nAdj="N", a=-2, alpha1=0.0, beta1=.2, alpha2=0.35);

Title " Sensitivity Analysis (with SSR)";
%DCSPEqNormal(Model="sum", alpha=0.05, beta=0.2, sigma=.3,
 NId=0.05, ux=0.70, uy=0.72, nInterim=600, Nmax=2000, N0=1200,
DuHa=0.01, nAdj="Y", a=-2, alpha1=0.0, beta1=.2, alpha2=0.35);
‹‹SAS‹‹
```

The simulation results are summarized as follows: under the null condition ($u_x = 0.70$, $u_y = 0.75$), the average sample-size is $\bar{N} = 868$/group and $\alpha = 0.050$; under the alternative condition ($u_x = 0.70$, $u_y = 0.71$), $\bar{N} = 1552$ and power = 0.91; if the treatment difference is bigger, e.g., 2% ($u_x = 0.70$, $u_y = 0.72$), the average sample-size $\bar{N} = 1088$ and power = 0.72 for the design without SSR, and $\bar{N} = 1515$ and power = 0.78 for the design with SSR. There is about 6% power increase by using SSR.

4.7 Summary

We have derived the closed forms for stopping boundaries using stagewise p-values based on subsamples from each stage, in which we have assumed the p-values are independent and uniformly distributed without proof or precisely the p-values are p-clud (See Chapter 8). The test statistics corresponding to the three methods (MIP, MSP, and MPP) are individual stagewise p-values without combining the data from different stages, the sum of the stagewise p-values, and the product of the stagewise p-values. Note that MIP is different from classic design because the early futility boundaries are used to construct the stopping boundaries for later stages. By comparing the results from MSP, MPP, or other methods with MIP, we can study how much efficacy is gained by combining data from different stages. It is strongly suggested that sufficient simulations are performed using SAS macros provided in this chapter or R programs in the appendix before determining which method should be used (the electronic versions of the simulation programs can be obtained at www.statisticians.org). There are practical issues that need to be considered too, which will be addressed in later chapters. Finally, although the examples of adaptive design in this chapter mainly involve sample-size re-estimation, MIP, MSP, and MPP are general adaptive design methods and can be used for a variety of adaptive designs (see later chapters of this book).

Problem

4.1 Suppose the median times for the two treatment groups in Example 4.7 are 9 months and 12 months. Design an adaptive trial and justify the adaptive design method (MIP, MSP, or MPP) and the design you have chosen.

4.2 Derive stopping boundaries and stagewise-ordering p-value for two-stage adaptive design based on the following test statistics:

$$T_k = \frac{1}{k}\sum_{i=1}^{k} p_i \text{ for k=1 and 2,}$$

where p_i is the stagewise p-value based on subsample from the i^{th} stage.

4.3 Investigate the independence of the stagewise test statistics (4.18) under exponential survival distribution and modify SAS Macro 4.4 to allow for sample-size re-estimation.

Chapter 5

Method with Inverse-Normal P-values

In this chapter we will study the method based on inverse-normal p-values (MINP), in which the test statistic at the k^{th} stage T_k is a linear combination of the weighted inverse-normal of the stagewise p-values. The weight can be fixed or a function of information time. MINP can be viewed as a general method including the group sequential method, the Lan-DeMets error-spending method, the Lehmacher-Wassmer method, the Fisher-Shen self-design method, and the Cui-Hung-Wang SSR method as special cases.

5.1 Method with Linear Combination of Z-Scores

Let z_k be the stagewise Normal test statistic at the k^{th} stage. In a group sequential design, the test statistic can be expressed as

$$T_k^* = \sum_{i=1}^{k} w_{ki} z_i, \tag{5.1}$$

where the weights satisfy the equality $\sum_{i=1}^{k} w_{ki}^2 = 1$ and the stagewise Normal statistic z_i is based on the subsample for the i^{th} stage. The weights w_{ki} can be functions of information time or sample-size fraction. Note that for fixed weights, T_k^* in (5.1) has a Normal distribution. For weights that are a function of information time, T_k^* in (5.1) forms a Brownian motion. Utilization of (5.1) with constant weights allows for changes in the timing (information time) of the interim analyses and the total sample-size, while the incorporation of functional weights allows for changes in timing and in the number of analyses. Their combination will allow broader adaptations.

Note that when w_{ki} is fixed, the standard multi-variate normal distribution of $\{T_1^*, ..., T_k^*\}$ will not change regardless of adaptations as long as z_i

$(i = 1, ..., k)$ has the standard normal distribution. To be consistent with the unified formations in Chapter 3, in which the test statistic is on p-scale, we use the transformation $T_k = 1 - \Phi\left(T_k^*\right)$ such that

$$T_k = 1 - \Phi\left(\sum_{i=1}^{k} w_{ki} z_i\right), \tag{5.2}$$

where Φ = c.d.f. of the standard normal distribution.

Unlike MIP, MSP, and MPP, the stopping boundary and power in MINP can be calculated using only numerical integration or computer simulation. Numerical integration is complicated (Jennison and Turnbull, 2000; CTriSoft, 2002), but the determination of stopping boundaries through simulation is straight forward and precise. The stopping boundaries based on the test statistic defined by (5.2) can be generated using SAS macro 5.1.

SAS Macro 5.1 is implemented for computing stopping boundaries with MINP. The SAS variables are defined as follows: **nSims** = number of simulations, **Model** = "fixedW" for constant weights; otherwise, the weights are dependent on information time. For the constant weights, values are specified by **w1** and **w2**. **alpha** = familywise α; **nInterim** = sample-size per group for the interim analysis; and **Nmax** = maximum sample-size allowed for the trial. **alpha1**, **beta1**, and **alpha2** are the stopping boundaries. **Power** is the probability of rejecting the null hypothesis. Therefore, the boundaries should be adjusted during the simulation until the power is equal to **alpha**.

>>SAS Macro 5.1: Stopping Boundaries with Adaptive Designs>>

```
%Macro SB2StgMINP(nSims=10000000, Model="fixedW", w1=0.5,
        w2=0.5, alpha=0.025, nInterim=50, Nmax=100,
        alpha1=0.01, beta1=0.15, alpha2=0.1871);
Data SB2StgMINP; Keep Model Power;
alpha=&alpha; Nmax=&Nmax; Model=&model;
 w1=&w1; w2=&w2; n1=&nInterim; Power=0; seedx=231;
 Do isim=1 To &nSims;
 nFinal=N1;
     T1 = Rannor(seedx);
    p1=1-ProbNorm(T1);
    If p1<=&alpha1 then do;
        Power=Power+1/&nSims;
    End;
    if p1>&alpha1 and p1<=&beta1 then do;
```

```
            T2 = Rannor(seedx);
             If Model^="fixedW" Then do
                 w1=Sqrt(n1/nFinal);
                 w2=Sqrt((1-n1/nFinal));
            End;
            Z2=(w1*T1+w2*T2)/Sqrt(w1*w1+w2*w2);
        p2=1-ProbNorm(Z2);
             If .<p2<=&alpha2 then Power=Power+1/&nSims;
        End;
    End;
    Output;
    Run;
    Proc Print data=SB2StgMINP; Run;
    %Mend SB2StgMINP;
    ‹‹SAS‹‹
```

To determine the stopping boundaries, prefix **alpha1** and **beta1**, and try different **alpha2** until **Power** from the SAS output is close enough to alpha. An example of calling SAS Macro 5.1 is presented in the following:

```
    ››SAS››
    Title "Stopping Boundaries by Simulations";
    %SB2StgMINP(Model="fixedW", w1=0.5, w2=0.5, alpha=0.025,
    nInterim=50, Nmax=100, alpha1=0, beta1=0.15, alpha2=0.0327);
    ‹‹SAS‹‹
```

Examples of stopping boundaries for a two-stage design with equal weights are presented in Table 5.1.

Table 5.1: Stopping Boundaries α_2 with Equal Weights

	α_1	0.000	0.0025	0.005	0.010	0.015	0.020
β_1							
0.15		0.0327	0.0315	0.0295	0.0244	0.0182	0.0105
0.20		0.0295	0.0284	0.0267	0.0221	0.0165	0.0097
0.25		0.0279	0.0267	0.0250	0.0209	0.0156	0.0092
0.30		0.0268	0.0257	0.0241	0.0202	0.0152	0.0090
0.35	α_2	0.0262	0.0251	0.0236	0.0197	0.0148	0.0089
0.50		0.0253	0.0243	0.0228	0.0191	0.0144	0.0087
0.70		0.0250	0.0240	0.0226	0.0189	0.0143	0.0086
1.00		0.0250	0.0240	0.0225	0.0188	0.0143	0.0086

Note: One-sided $\alpha = 0.025$ and 10,000,000 simulation runs

5.2 Lehmacher and Wassmer Method

To extend the method with linear combination of z-scores, Lehmacher and Wassmer (1999) proposed the test statistic at the k^{th} stage that results from the inverse-normal method of combining independent stagewise p-values (Hedges and Olkin, 1985):

$$T_k^* = \sum_{i=1}^{k} w_{ki} \Phi^{-1}(1 - p_i), \tag{5.3}$$

where the weights satisfy the equality $\sum_{i=1}^{k} w_{ki}^2 = 1$, and Φ^{-1} is the inverse function of Φ, the standard normal c.d.f. Under the null hypothesis, the stagewise p_i is usually uniformly distributed over [0,1]. The random variables $z_i = \Phi^{-1}(1 - p_i)$ and T_k^* have the standard normal distribution. Lehmacher and Wassmer (1999) suggested using equal weights, i.e., $w_{ik} \equiv \frac{1}{\sqrt{k}}$.

Again, to be consistent with the unified formulations proposed in Chapter 3, transform the test statistic to the p-scale, i.e.,

$$T_k = 1 - \Phi\left(\sum_{i=1}^{k} w_{ki} \Phi^{-1}(1 - p_i)\right). \tag{5.4}$$

With (5.4), the stopping boundary is on the p-scale and easy to compare with other methods regarding operating characteristics. In this book, (5.4) is implemented in SAS and R, instead of (5.3).

When the test statistic defined by (5.4) is used, the classical group sequential boundaries are valid regardless of the timing and sample-size adjustment that may be based on the observed data at the previous stages. Note that under the null hypothesis, p_i is usually uniformly distributed over [0,1] and hence $z_i = \Phi^{-1}(1 - p_i)$ has the standard normal distribution; so does T_k. The Lehmacher-Wassmer method provides a broad method for different endpoints as long as the p-value under the null hypothesis is uniformly distributed over [0,1].

The Lehmacher-Wassmer Inverse-Normal method has been implemented in SAS and R (SAS Macro 5.2). The SAS variables are defined as follows: **nSims** = number of simulations, **Model** = "fixedW" for constant weights, and **Model** = "InfoFun" for functional weights, **alpha** = overall alpha level, **sigma** = the equivalent standard deviation, **NId** = the noninferiority margin, and **ux** and **uy** = the responses in groups x and y, respectively. **nInterim** = sample-size for the interim analysis, **Nmax** = maximum sample-size, **N0** = initial sample-size at the final analysis, and

DuHa = the treatment difference under alternative hypothesis. **nAdj** = "Y" for sample-size adjustment; for no SSR, **nAdj** = "N." **a** = the parameter in the sample-size adjustment algorithm (4.17) or (9.1), and **alpha1**, **beta1**, and **alpha2** = stopping boundaries as defined before. **FSP** = futility stopping probability, **ESP** = efficacy stopping probability, **AveN** = average sample-size, **Power** = power from simulations, and **nClassic** = sample-size for the corresponding classic design.

››SAS Macro 5.2: Two-Stage Design with Inverse-Normal Method››

```
%Macro MINP(nSims=1000000, Model="fixedW", w1=0.5,
    w2=0.5, alpha=0.025, beta=0.2, sigma=2, NId=0, ux=0, uy=1,
    nInterim=50, Nmax=100, N0=100, DuHa=1, nAdj="Y", a=2,
    alpha1=0.01, beta1=0.15, alpha2=0.1871);
Data MINP; Keep Model FSP ESP AveN Power nClassic PAdj;
 seedx=1736; seedy=6214; alpha=&alpha; NId=&NId;
Nmax=&Nmax; Model=&model; w1=&w1; w2=&w2; ux=&ux;
uy=&uy; sigma=&sigma;  N1=&nInterim;
eSize=abs(&DuHa+NId)/sigma;
 nClassic=Round(2*((Probit(1-alpha)+Probit(1-&beta))/eSize)**2);
 FSP=0; ESP=0; AveN=0; Power=0;
 Do isim=1 To &nSims;
 nFinal=N1;
    ux1 = Rannor(seedx)*sigma/Sqrt(N1)+ux;
    uy1 = Rannor(seedy)*sigma/Sqrt(N1)+uy;
    T1 = (uy1-ux1+NId)*Sqrt(N1)/2**0.5/sigma;
    p1=1-ProbNorm(T1);
    If p1>&beta1 Then FSP=FSP+1/&nSims;
    If p1<=&alpha1 Then Do;
        Power=Power+1/&nSims; ESP=ESP+1/&nSims;
    End;
    If p1>&alpha1 and p1<=&beta1 Then Do;
        eRatio=Abs(&DuHa/(Abs(uy1-ux1)+0.0000001));
        nFinal=min(&Nmax,max(&N0,eRatio**&a*&N0));
        If &DuHa*(uy1-ux1+NId)<0 Then nFinal=N1;
        If &nAdj="N" Then nFinal=&Nmax;
        If nFinal>N1 Then Do;
            ux2 = Rannor(seedx)*sigma/Sqrt(nFinal-N1)+ux ;
            uy2 = Rannor(seedy)*sigma/Sqrt(nFinal-N1)+uy;
            T2 = (uy2-ux2+NId)*Sqrt(nFinal-N1)/2**0.5/sigma;
            If Model^="fixedW" Then Do
```

```
            w1=Sqrt(N1/nFinal);
            w2=Sqrt((1-N1/nFinal));
          End;
          Z2=(w1*T1+w2*T2)/Sqrt(w1*w1+w2*w2);
          p2=1-ProbNorm(Z2);
          If .<p2<=&alpha2 Then Power=Power+1/&nSims;
        End;
      End;
      AveN=AveN+nFinal/&nSims;
  End;
  PAdj=&alpha1+power-ESP;  ** Stagewise ordering p-value;
  Output;
  Run;
  Proc Print Data=MINP; Run;
  %Mend MINP;
  ‹‹SAS‹‹
```

Example 5.1 Inverse-Normal Method with Normal Endpoint

Let's use an earlier example of an asthma study. Suppose a phase-III asthma study with 2 dose groups (control and active) with the percent change from baseline in FEV1 as the primary efficacy endpoint. The estimated FEV1 improvement from baseline is 5% and 12% for the control and active groups, respectively, with a common standard deviation of $\sigma = 22\%$. Based on a large sample assumption, the sample-size for a fixed design is 208 per group with 90% power and a one-sided alpha = 0.025. Using MIP, an interim analysis is planned based on the response assessment for 50% of the patients. We now use SAS Macro 5.2 to assist in the adaptive design, described as follows:

(1) Choose stopping boundaries at the first stage: $\alpha_1 = 0.01$, $\beta_1 = 1$; then from Table 5.1, we obtain the corresponding $\alpha_2 = 0.019$.

(2) Check the stopping boundary to make sure that the familywise error is controlled by using the following SAS statement:

```
  ››SAS››
  %MINP(Model="fixedW", w1=0.5, w2=0.5, alpha=0.025, beta=0.1,
  sigma=0.22, ux=0.05, uy=0.05, nInterim=100, Nmax=200, DuHa=0.07,
  nAdj="N", alpha1=0.01, beta1=1, alpha2=0.019);
  ‹‹SAS‹‹
```

The simulated familywise error rate $\alpha = 0.0253$. Therefore the stopping boundaries are confirmed.

(3) Calculate power or sample-size required using the following SAS statement:

```
>>SAS>>
%MINP(Model="fixedW", w1=0.5, w2=0.5, alpha=0.025, beta=0.1,
sigma=0.22, ux=0.05, uy=0.12, nInterim=100, Nmax=200, DuHa=0.07,
nAdj="N", alpha1=0.01, beta1=1, alpha2=0.019);
<<SAS<<
```

(4) Perform the sensitivity analysis under the condition H_s: 0.05 versus 0.1 by submitting the following SAS code:

```
>>SAS>>
%MINP(Model="fixedW", w1=0.5, w2=0.5, alpha=0.025, beta=0.1,
sigma=0.22, ux=0.05, uy=0.10, nInterim=100, Nmax=200, DuHa=0.07,
nAdj="N", alpha1=0.01, beta1=1, alpha2=0.019);
<<SAS<<
```

We now summarize the simulation outputs of the three scenarios in Table 5.2.

Table 5.2: Operating Characteristics of GSD with MSP

Simulation condition	FSP	ESP	$\bar{\text{N}}$	N_{max}	Power (alpha)
H_o	0	0.010	199	200	(0.025)
H_a	0	0.470	153	200	0.873
H_s	0	0.237	176	200	0.597

Note that OBF boundary is for early efficacy stopping only and the corresponding error spending for a one-sided test is $\alpha_1 = 0.025$, $\beta_1 = 1$, and $\alpha_2 = 0.0240$.

We now calculate stagewise-ordering adjusted p-values (See Chapter 3). If the trial stopped at the first stage, then the p-value does not need any adjustment. Suppose the trial is finished, with a stagewise p-value for the first stage of $p_1 = 0.012$ (which is larger than $\alpha_1 = 0.01$ and not significant; therefore the trial continued to the second stage) and the stagewise p-value for the second stage of $p_2 = 0.015 < \alpha_2 = 0.019$. Therefore the null hypothesis is rejected. However, $p_2 = 0.019$ is the naive or unadjusted p-value. The stagewise-ordering adjusted p-value at stage 2 can be obtained through simulation, which is illustrated as follows:

The conditional (on trial stopping at stage 2) p-value is the probability of the stagewise p-value at stage 2 being smaller than the observed

stagewise p-value. Therefore, we can use the same SAS Macro 5.2 for the power calculation to calculate the conditional p-value. To do this, we use the observed stagewise p-value p_2 to replace α_2 in SAS Macro 5.2; then stagewise-ordering p-value is $p_{adj} = \alpha_1 + p_c$ from the SAS output. The following is the SAS call to generate (under Ho) the adjusted p-value:

```
>>SAS>>
%MINP(Model="fixedW", w1=0.5, w2=0.5, alpha=0.025, beta=0.1,
sigma=0.22, ux=0.05, uy=0.05, nInterim=155, Nmax=310, DuHa=0.07,
nAdj="N", alpha1=0.01, beta1=1, alpha2=0.015);
<<SAS<<
```

From the SAS outputs, the stagewise-ordering p-value is $p_{adj} = 0.0215 < \alpha$.

Example 5.2 Inverse-Normal Method with SSR

Now suppose we want to do sample-size re-estimation (SSR) and the interim analysis is planed for 100 patients/group. The SSR rule is given by (4.17) with the parameter of a = 2. We present two designs with equal weights: (1) The trial does not allow for early stopping and the interim analysis is for sample-size re-estimation only. The stopping boundaries are $\alpha_1 = 0$, $\beta_1 = 1$, and $\alpha_2 = 0.025$; and (2) The interim analysis is for early futility stopping and sample-size re-estimation. The stopping boundaries are $\alpha_1 = 0$, $\beta_1 = 0.5$, and $\alpha_2 = 0.0253$. For SSR, the maximum sample-size is $N_{max} = 400$/group, and the initial sample-size is No = 200/group. We are going to assess the n-re-estimation mechanism when the treatment difference is small (5% versus 10%) using SAS Macro 5.2.

For design 1, the simulations can be performed using the following SAS macro call:

```
>>SAS>>
%MINP(Model="fixedW", w1=1, w2=1, alpha=0.025, beta=0.1,
sigma=0.22, ux=0.05, uy=0.1, nInterim=100, Nmax=400, N0=200,
nAdj="Y", DuHa=0.07, alpha1=0, beta1=1, alpha2=0.025);
<<SAS<<
```

For design 2, the simulations can be performed using the following SAS macro call:

```
>>SAS>>
%MINP(Model="fixedW", w1=1, w2=1, alpha=0.025, beta=0.1,
sigma=0.22, ux=0.05, uy=0.1, nInterim=100, Nmax=400, N0=200,
```

nAdj="Y", DuHa=0.07, alpha1=0, beta1=0.5, alpha2=0.0253);
‹‹**SAS**‹‹

We now summarize the simulation results in Table 5.3.

Table 5.3: Operating Characteristics of a SSR with MSP

Design	FSP	ESP	$\bar{\text{N}}$	N_{max}	Power
Without SSR	0.167	0.237	176	200	0.597
SSR only	0	0	304	400	0.823
SSR. & futility stopping	0.054	0	304	400	0.825

Note that the results for the design without SSR (i.e., classic group sequential design) are from the sensitivity analysis in Table 5.2. From Table 5.3, we can see that there are similar operating characteristics between the two designs with SSR. Both designs increase the power from 59.7 for the classic group sequential design to over 82%.

5.3 Classic Group Sequential Method

In classic group sequential design (GSD), the stopping boundaries are usually specified by a function of stage k. The commonly used such functions are Pocock and O'Brien-Fleming boundary functions. Wang and Tsiatis (1987) proposed a family of two-sided tests with a shape parameter Δ, which includes Pocock's and O'Brien-Fleming's boundary functions as special cases. Because W-T boundary is based on z-scale, for consistent, we can convert them to p-scale. The W-T boundary on p-scale is given by

$$a_k > 1 - \Phi\left(\alpha_K \, I_k^{\Delta - 1/2}\right), \tag{5.5}$$

where $I_k = \frac{k}{K}$ or $I_k = \frac{n_k}{n_K}$ (information time), α_K is the stopping boundary at the final stage and a function of the number of stages K, α,and Δ.

Note that the Normal statistic T_k from the cumulative sample at the k^{th} stage in GSD can be written in the combination of stagewise Normal z-statistics $(z_i, i = 1, ..., k)$:

$$T_k = \frac{1}{\sqrt{N_k}} \sum_{j=1}^{N_k} (y_j - x_j) = \frac{1}{\sqrt{N_k}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{i_j} - x_{i_j}\right)$$
$$= \sum_{i=1}^{k} \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \left(y_{i_j} - x_{i_j}\right) \sqrt{\frac{n_i}{N_k}} = \sum_{i=1}^{k} z_i \sqrt{\eta_{ki}},$$

where x_i and y_i are the i^{th} observations in group x and y, respectively, n_i = stagewise sample-size at stage i, $N_k = \Sigma_{j=1}^{k} n_j$ is the cumulative sample-size at stage k, and the information fraction (not information time!) $\eta_{ki} = \frac{n_i}{N_k} = w_{ki}^2$, and $z_i = \frac{1}{\sqrt{n_i}} \sum_{j=1}^{n_i} \left(y_{i_j} - x_{i_j}\right)$. For simplicity, we have assumed $\sigma = 1$ in the derivation.

Therefore, in classic GSD method, we can write the test statistic at the k^{th} stage as

$$T_k = \sum_{i=1}^{k} w_{ki} z_i, \tag{5.6}$$

where the weights

$$w_{ki} = \sqrt{\eta_{ki}}. \tag{5.7}$$

The method can be extended to other endpoint using the inverse-normal transform, i.e.,

$$T_k = \sum_{i=1}^{k} \sqrt{\eta_{ki}} \Phi^{-1} (1 - p_i). \tag{5.8}$$

For a two-stage design with two independent groups, (5.6) becomes

$$\begin{cases} T_1 = z_1 \\ T_2 = z_1 \sqrt{\frac{n_1}{n_1+n_2}} + z_2 \sqrt{\frac{n_2}{n_1+n_2}}. \end{cases} \tag{5.9}$$

For a group sequential design without SSR the weights $w_{ki} = \sqrt{\eta_{ki}}$ is a prefixed constant, which is basically the same as the L-W method from the previous section. However, it allows for SSR; $\sqrt{\eta_{ki}} (i > 1)$ is a function of $z_j (j = 1, ..., i-1)$. Therefore, T_k is not a linear combination of z_i; hence, it is usually not Normal. Consequently, the stopping boundaries for the classic group sequential designs cannot be used in the case with sample-size re-estimation. In other words, when the test statistic is defined by (5.6),

a new set of stopping boundaries has to be determined using computer simulation when the trial allows for SSR.

The following are the numerical examples of this method used with and without SSR.

Example 5.3 Group Sequential Design

We will use the asthma trial example again. A phase III asthma study with 2 dose groups (control and active) with the percent change from baseline in FEV1 as the primary efficacy endpoint. The estimated FEV1 improvement from baseline is 5% and 12% for the control and active groups, respectively, with a common standard deviation of $\sigma = 22\%$. The interim analysis is performed based on the first 100 patients in each group. A futility design is used with $\alpha_1 = 0, \beta_1 = 0.5$, and $\alpha_2 = 0.0253$.

In simulations with the SAS Macro 5.2, the model is specified as "GSD" (group sequential design). Again, we study the sensitivity by assuming 5% versus 10% FEV1 change in the control and test groups, respectively. (Note that DuHa = 0.07 not 0.05 should be used for the sensitivity analysis)

(1) Design without sample-size adjustment

Because this is not equal weights design and the interim analysis was not performed for 50% of the patients, we cannot use the stopping boundaries in Table 5.1. The stopping boundaries can be determined using simulations. By fixing $\alpha_1 = 0$, $\beta_1 = 0.5$, and trying different values for α_2, we find that $\alpha_1 = 0$, $\beta_1 = 0.5$, and $\alpha_2 = 0.0266$ satisfy the requirement of overall alpha = 0.025. Here is the final SAS macro call to obtain the stopping boundaries:

```
>>SAS>>
%MINP(Model="GSD", alpha=0.025, beta=0.1, sigma=0.22,
ux=0.05, uy=0.05, nInterim=100, Nmax=310, DuHa=0.07,
nAdj="N", alpha1=0, beta1=0.5, alpha2=0.0266);
<<SAS<<
```

The power of the design can be obtained by using the following SAS statement:

```
>>SAS>>
%MINP(Model="GSD", alpha=0.025, beta=0.1, sigma=0.22,
ux=0.05, uy=0.10, nInterim=100, Nmax=310, DuHa=0.07,
nAdj="N", alpha1=0, beta1=0.5, alpha2=0.0266);
<<SAS<<
```

(2) Design with sample-size adjustment

For the design with sample-size adjustment, we have to first determine the stopping boundaries using simulations. By fixing $\alpha_1 = 0$, $\beta_1 = 0.5$, and

trying different α_2, we find that $\alpha_1 = 0$, $\beta_1 = 0.5$, and $\alpha_2 = 0.0265$ satisfy the overall alpha = 0.025 requirement. We use the following SAS code to determine the stopping boundaries and obtain the power, respectively.

```
>>SAS>>
%MINP(Model="GSD", alpha=0.025, beta=0.1, sigma=0.22, ux=0.05,
uy=0.05, nInterim=100, Nmax=400, N0=310, DuHa=0.07, alpha1=0,
nAdj="Y", beta1=0.5, alpha2=0.0265);

%MINP(Model="GSD", alpha=0.025, beta=0.1, sigma=0.22, ux=0.05,
uy=0.10, nInterim=100, Nmax=400, N0=310, DuHa=0.07,
nAdj="Y", alpha1=0, beta1=0.5, alpha2=0.0265);
<<SAS<<
```

The simulation results are presented in Table 5.4.

Table 5.4: Operating Characteristics of a GSD with MSP

Design	FSP	ESP	$\bar{N}$	N_{max}	Power
Group sequential	0.054	0	299	310	0.795
Adaptive SSR	0.054	0	356	400	0.861

5.4 Cui-Hung-Wang Method

Cui, Hung, and Wang (1999) developed a method for an adaptive design allowing for sample-size re-estimation based on the unblinded results of the interim analysis. Consider a group sequential trial with two groups and one interim analysis.

$$T_2^* = w_1 z_1 + w_2 z_2, \tag{5.10}$$

where the weights $w_i = \sqrt{\frac{n_i}{n_1+n_2}}$, in which the sample-size n_1 and n_2 are the original sample sizes. Because the original n_2 can be arbitrarily chosen, the weight $w_i > 0$ can actually be any prefixed positive value satisfying $w_1^2 + w_2^2 = 1$. It is important to remember that the weights are dependent on the originally planned sample-size not the modified sample-size. When there is actually no modification of sample-size the test statistic is the same as for the classic group sequential design.

There are many possible sample-size adjustment algorithms. Cui, et al. (1999) suggest using the following formulation for new sample-size at the

second stage:

$$n_2^* = \left(\frac{\delta}{\hat{\delta}_1}\right)^2 (n_1 + n_2) - n_1, \tag{5.11}$$

where δ and $\hat{\delta}$ are the initial estimated treatment difference and the observed difference at stage 1, respectively. The stopping boundaries for Ciu-Hung-Wang's method are the same as for a classic GSD.

5.5 Lan-DeMets Method

The Lan-DeMets method (Lan and DeMets, 1983) is an early and very interesting adaptive design method, the error-spending approach, in which they elegantly use the properties of Brownian (Wiener) motion.

5.5.1 *Brownian Motion*

Brownian motion (Figure 5.1) has been widely studied and the results are ready to use for sequential designs in clinical trials.

Definition 5.1: A stochastic process $\{X(t), t \geq 0\}$ is said to be a Brownian motion with a Drift μ if

(1) $X(0) = 0$;

(2) $\{X(t), t \geq 0\}$ has a stationary and independent increment;

(3) for every $t > 0, X(t)$ is normally distributed with mean μt and variance $\sigma^2 t$, i.e.,

$$X(t) \sim \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{(x-\mu t)^2}{2\sigma^2 t}\right). \tag{5.12}$$

.

The covariance of the Brownian motion is $\text{cov}[X(t), X(s)] = \sigma^2 \min\{s, t\}$.

The standard Brownian motion $B(t)$ is the Brownian motion with $\mu = 0$ and $\sigma^2 = 1$. The conditional probability density function of $B(t)$ is given by

$$p(x_2, t|x_1) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{1}{2t}(x_2 - x_1)^2\right), \tag{5.13}$$

where $x_1 = x_1(t_1)$, $x_2 = x_2(t_2)$, and $t_2 > t_1$.

The conditional probability of the position at time $t + s$ given the position at time s can be written as

$$\Pr\left\{B\left(t+s\right)\leq y|B\left(s\right)=x\right\}=\Phi\left(\frac{y-x}{\sqrt{t}}\right). \tag{5.14}$$

Because of the independent increment, the joint probability distribution of $X\left(t_1\right)...X\left(t_n\right)$ is given by

$$f\left(x_1...x_n\right)=\sum_{i=1}^{n}f_{t_i-t_{i-1}}\left(x_i-x_{i-1}\right)=\frac{\exp\left(-\frac{1}{2}\Sigma_{i=1}^{n}\frac{\left(x_i-x_{i-1}\right)^2}{t_i-t_{i-1}}\right)}{\left(2\pi\right)^{n/2}\sqrt{\Pi_{i=1}^{n}\left(t_i-t_{i-1}\right)}}. \tag{5.15}$$

The relationship between the standard Brownian motion and Brownian motion with drift μ can be expressed as

$$X\left(t\right)=\mu t+\sigma B\left(t\right). \tag{5.16}$$

The cumulative probability is given by

$$\Pr\left\{X\left(t\right)\leq y|X\left(0\right)=x\right\}=\Phi\left(\frac{y-x-\mu t}{\sigma\sqrt{t}}\right). \tag{5.17}$$

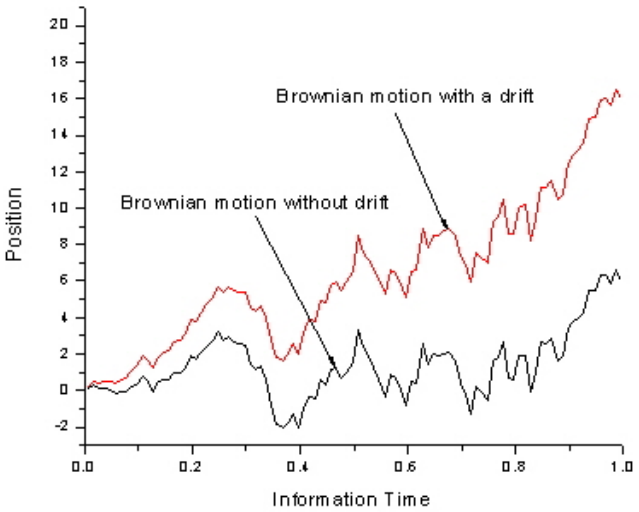


Figure 5.1: Examples of Brownian Motion

The First Hitting of Standard Brownian Motion:

Let C be a horizontal boundary (Figure 5.1) and $M(t)$ be the maximum of the standard Brownian motion with time t, i.e., $M(t) = \max_{0 \le u \le t} B(u)$. It can be proved, using the reflection principle (Taylor and Karlin, 1998, p.491-493), that the probability of the first passing (the boundary) before time $t > 0$ can be expressed as:

$$\Pr\{M(t) \ge C\} = 2\left[1 - \Phi\left(\frac{C}{\sqrt{t}}\right)\right]. \tag{5.18}$$

Equation (5.18) can be used directly to control type-I error (see next section).

Remarks: There are many examples of Brownian motion: Einstein showed that the solution ν for the diffusion or permeability equation $\frac{\partial \nu}{\partial t} = \frac{1}{2}\sigma^2 \frac{\partial^2 \nu}{\partial x^2}$ is the Brownian motion (5.12) with a unit diffusion coefficient. Random-walk with a varied step length forms a Brownian motion. Brownian motion is also known as a memoryless process. The independence of the increment process also implies that we can predict the future as long as we know the current status. However, in many cases, to predict the future, we have to know not only the present, but also the past.

5.5.2 *Lan-DeMets Error-Spending Method*

Brownian motion was first introduced by Lan and DeMets (1983) to the adaptive design with a prefixed error spending function, which allows for changing the timing and the number of analyses.

We know from (5.6) and (5.7) that the test statistic based on the cumulative sample-size at the k^{th} stage can be written as

$$Z_k = \Sigma_{i=1}^{k} z_i \sqrt{\eta_{ki}}. \tag{5.19}$$

When the maximum sample-size N is fixed, i.e., without SSR, the Brownian motion can be constructed as follows:

$$B_k = Z_k \sqrt{I_k}, \tag{5.20}$$

where the information time $I_k = \frac{N_k}{N}$, $N_k = \Sigma_{i=1}^{k} n_i$.

From (5.20), the following properties can be obtained using simple calculations:

(1) $E[B_N(I)] = \theta\sqrt{N}$.

(2) $var(B_N(I)) = I$,

(3) $cov(B_N(I_1), B_N(I_2)) = \min(I_1, I_2)$

Note that B_k is a linear function of information time $I_k \in [0,1]$. Because the Brownian motion is not observable between two interim analyses, we can assign an accumulated crossing probability to the information time point I_k.

Brownian motion is formed only if the trial continues without any early stopping. However, if we are interested in the first pass (efficacy or futility), then Brownian motion results can be used for the trial with early stopping.

We can see that the Brownian motion can be viewed as weighted stage-wise z-scores, where the weights are

$$w_{ki} = \sqrt{I_k \eta_{ki}}. \tag{5.21}$$

The Lan-DeMets method is similar to, but different from the L-W method because the weights $w_{ki} = \sqrt{I_k \eta_{ki}}$ is not a prefixed constant. Instead, it is a prefixed function of information time. Note that the Lan-DeMets method uses the same stopping boundaries as a classic GSD, because for each fixed information time, the test statistic is the same as a classic GSD. For the two-stage design, the stopping boundaries and power can be obtained through simulation using SAS Macros 5.1 and 5.2.

We now use Brownian motion to illustrate the error-spending method because Lan and DeMets (1983) originally proposed the error-spending approach using Brownian motion, although the error-spending can be used with other test statistics (see Chapter 6).

If H_o is rejected whenever the position of the Brownian motion first crosses the boundary C, then we can control overall α by letting the maximum crossing probability $\Pr\{M(1) \geq C\} = \alpha$, and solving (5.18) for C. In other words, from $2[1 - \Phi(C)] = \alpha$, we can immediately obtain $C = z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$. We now designate the error-spending function to be the first passing probability (5.18), i.e.,

$$\alpha^*(I_k) = \begin{cases} 2\left[1 - \Phi\left(\frac{z_{1-\alpha/2}}{\sqrt{I_k}}\right)\right], & I_k > 0 \\ 0, & I_k = 0 \end{cases}. \tag{5.22}$$

Note that $\alpha^*(t)$ is a increasing function in t or information time I_k and $\alpha^*(1) = \alpha$, the one-sided significance level.

As stated in Chapter 3, for the error-spending approach, the stopping boundaries are determined by

$$\Pr\left\{\cap_{j=1}^{k-1}\left[\beta_j < B(t_j) < \alpha_j\right] \cap B(t_k) \geq \alpha_k\right\} = \alpha^*(I_k) - \alpha^*(I_{k-1}). \tag{5.23}$$

Using (5.15), (5.23) becomes

$$\int_{\beta_1}^{\alpha_1} \cdots \int_{\beta_{i-1}}^{\alpha_{k-1}} \int_{\alpha_k}^{\infty} \frac{\exp\left(-\frac{1}{2}\Sigma_{i=1}^{n} \frac{(x_i - x_{i-1})^2}{t_i - t_{i-1}}\right)}{(2\pi)^{n/2} \sqrt{\Pi_{i=1}^{n} (t_i - t_{i-1})}} dx_1 \cdots dx_{k-1} dx_k$$
$$= \alpha^* (I_k) - \alpha^* (I_{k-1}) . \tag{5.24}$$

Lan and DeMets (1983) formulated the problem without early futility stopping boundaries. Here it is generalized to the design allowing for futility stopping. Determination of the stopping boundaries $(\beta_i, \alpha_i; i = 1, ..., .K)$ requires either numerical integration (Armitage et al., 1969; Jennison and Turnbull, 2000) or computer simulations (See Chapter 6).

The error-spending function can be any non-decreased error spending function $\alpha^* (t)$ with a range of [0,1]. When $\alpha^* (I_k) = \alpha^* (I_{k-1})$, the k^{th} stage interim analysis is used either for futility stopping or modifying the design (such as its randomization), but not for efficacy stopping. Note that (5.22) is the error-spending function corresponding to the O'Brien-Fleming stopping boundaries. Other commonly used error spending functions include Pocock's $\alpha^*(t) = \alpha \log[1 + (e - 1)t]$ (Kim and DeMets,1992), and power family: $\alpha^*(t) = \alpha t^\theta, \theta > 0$.

When the error-spending function $\alpha^*(t)$ is prefixed, i.e., not dependent on the observed data from the trial, then the overall type-I error rate is $\Sigma_{k=1}^{K} \pi_k = \Sigma_{k=1}^{K} [\alpha^* (I_k) - \alpha^* (I_{k-1})] = \alpha^* (1) - \alpha^* (0) = \alpha$. This is true even when the number of analyses K and the timing of the analyses I_k are not predetermined. This is the most attractive feature of the error-spending function. We further illustrate the approach in the following example.

Example 5.4 Changes in Number and Timing of Interim Analyses

An international, multi-center, randomized phase-III study to compare the test drug with a combination of drugs in patients with newly diagnosed multiple myeloma was designed using the O'Brien-Fleming spending function. The interim analysis was to be performed for 200 patients. The final analysis will be performed using 400 patients. The primary study objective is to assess the treatment difference in overall complete response rate (CR) obtained at the end of a 16-week induction phase. However, due to the complexity of the international trial, the data collection and validation became extremely challenging. It was decided that the investigator's assessment would be used because it is available much earlier than the assessment by the independent review committee (IRC) — the gold standard. However, the discrepancies between the two assessments are not

known. The sponsor is concerned that if the trial is stopped based on the IDMC's recommendation, which is based on the investigator's assessment, it could be found later that the treatment difference is not significant based on the IRC's assessment. However, it is known that when the trial is stopped at the first interim analysis (IA), there will be more patients enrolled (about 300). Therefore, the sponsor decided to add a second interim analysis. The second IA is very interesting to the sponsor because if the results are significant at the first interim analysis based on investigator's assessment ($\alpha = 0.0025$), then the p-value based on the IRC's assessment should be somewhat close to 0.0025. With 300 patients (the exact number is based on the number of patients randomized at the first IA) at the second IA and based on the OB-F spending function (5.22), the error spent on the three analyses is $\pi_1 = 2\left[1 - \Phi\left(\frac{2.240}{\sqrt{200/400}}\right)\right] = 0.0016$, $\pi_2 = 2\left[1 - \Phi\left(\frac{2.240}{\sqrt{300/400}}\right)\right] - 0.0016 = 0.0096 - 0.0016 = 0.008$, and $\pi_3 = 2\left[1 - \Phi\left(2.240\right)\right] - 0.0096 = 0.025 - 0.0096 = 0.015\,4$. The actual π_2 should be based on the actual number of patients at the second IA.

5.6 Fisher-Shen Method

Fisher and Shen (Fisher, 1998; Shen and Fisher, 1999) propose a self-designing approach for k-stage designs. In this method, the test statistic is defined similarly to Lehmacher-Wassmer's method, i.e., it is the weighted sums of the standardized difference z_i. However, the weights w_i at each stage may be determined based on data from previous stages, and the number of stages K does not have to be prefixed, but the condition $\Sigma_{i=1}^{K} w_i^2 = 1$ must be met. Fisher (1998) does not consider early stopping in his method. Shen and Fisher (1999) consider early futility stopping ($\alpha_1 = 0$), but do not account for its impact on type-I error; hence, it is a conservative approach.

5.7 Summary

In this chapter we have studied broad methods that are based on weighted inverse-normal stagewise p-values. The weights can be fixed such as in the GSD, the L-W method, and the Cui-Hung-Wang method, or varied depending on observed data, like in the Fisher-Shen and Lan-DeMets methods. The Cui-Hung-Wang method can be used for sample-size re-estimation for a normal endpoint, and the L-W method can be used for SSR for various endpoints. The Lan-DeMets method can be used for adaptive designs

with changes in the number and the timing of analyses. Fisher-Shen is a method that can be used for SSR, but it is conservative because the futility boundaries at earlier stages are not used in the construction of later stopping boundaries. Deciding which method and design are best for a trial is heavily dependent on the practical setting. Simulations should be used to assist in decision-making. The SAS macros in this chapter provide powerful tools for accomplishing this goal. The electronic versions of the simulation programs can be obtained at www.statisticians.org.

Problem

5.1 Suppose the median times for the two treatment groups in Example 4.7 are 9 months and 12 months. Design an adaptive trial and justify the adaptive design method (MIP, MSP, MPP, MINP) and the design you have chosen.

5.2 Use the reflection principle (Taylor and Karlin, 1998, p.497) to prove that

$$\Pr\{M(t) \geq z,\ B(t) \leq y\} = \Pr\{B(t) \geq 2z - x\} = 1 - \Phi\left(\frac{2z-y}{\sqrt{t}}\right).$$

5.3 The Gamler's Ruin Problem (Taylor and Karlin,1998, p.509)

Theorem: For a Brownian motion with drift parameter μ and variance σ^2, and $a < x < b$,

$$u(x) = \Pr\{X(T_{ab}) = b | X(0) = x\} = \frac{e^{-2\mu x/\sigma^2} - e^{-2\mu a/\sigma^2}}{e^{-2\mu b/\sigma^2} - e^{-2\mu a/\sigma^2}}, \qquad (5.25)$$

where T_{ab} is a random time at which the process $X(t)$ first assumes one of the values a or b. The so-called hitting time is defined by

$$T_{ab} = \min\{t \geq 0; X(t) = a \text{ or } X(t) = b\}. \qquad (5.26)$$

It can be seen that $u(x)$ is the conditional probability of hitting threshold b given the first hit occurs.

The expectation of $u(x)$ is given by

$$E[T_{ab}|X(0) = x] = \frac{1}{\mu}[u(x)(b-a) - (x-a)].$$

It is interesting to know that we cannot obtain the velocity of the Brownian motion particle by taking the derivative of $B(t)$ with respect to t because $B(t)$ is continuous but not differentiable at any single point (Taylor and Karlin, 1998, p.509). This counter-intuitive fact is difficult to comprehend.

For $\mu -> 0$, and $\sigma -> 1$, (5.25) becomes

$$u(x) = \Pr\{B(T_{ab}) = b | B(0) = x\} = \frac{x-a}{b-a}.$$

Study the possibility of using these results in adaptive design.

Chapter 6

Implementation of K-Stage Adaptive Designs

6.1 Introduction

We are going to present simulation approaches to the N-stage design using nonparameteric stopping boundaries and the error spending approach in this chapter. The latter allows for modifying the timing and number of analyses. The two methods have been implemented in SAS with MIP, MSP, MPP, and MINP. We will illustrate how to use these programs to design adaptive trials.

In a nonparametric approach, stopping boundaries are determined by the overall α level without specification of any function for $\alpha\{i\}$ or error-spending function. Therefore this method may not be applicable to adaptive designs with changes in the number or timing of the analyses. To allow for changes in the number and timing of interim analyses, we can prespecify discretely when the possible times are and how much error is to spend at each interim analysis. In general, the stopping boundaries and power of an n-stage design can be determined using simulation or numerical integration regardless of the test statistic. Simulation is usually easier and computationally faster. Numerical integration usually requires dimension reductions in order to be computationally feasible. For n-stage group sequential designs, a numerical integration algorithm can be found in Jennison and Turbull (2000) and CTriSoft (2002). For general adaptive designs, numerical algorithms are not available.

6.2 Nonparametric Approach

6.2.1 *Normal Endpoint*

SAS macro 6.1 can be used for simulating two-arm N-stage adaptive designs with Normal endpoint. The SAS variables are defined as follows: **nSims** =

number of simulation runs, and **Model** = adaptive design method: "MIP," "MSP," or "MPP." **nStgs** = number of stages, **ux**, **uy** = means for groups x and y, respectively, **NId** = noninferiority margin, and **sigma** = standard deviation. **nAdj** = "Y" to allow for sample-size adjustment; otherwise, **nAdj** = "N." **N0** = initial cumulative sample-size for the final stage, $\sum_i$**Ns{i}**, and **Nmax** = maximum sample-size allowed. **DuHa** = true treatment mean difference; for power calculation, **DuHa** = **uy - ux**, and for sensitivity analysis, **DuHa** $\neq$ **uy** – **ux**. **Ns{i}** = the i^{th} stage sample-size (not cumulative one), **alpha{i}** = the efficacy stopping boundary at the i^{th} stage, and **beta{i}** = the futility stopping boundary at the i^{th} stage. **ESP{i}** = the efficacy stopping probability at the i^{th} stage, and **FSP{i}** = the futility stopping probability at the i^{th} stage. **power** = the simulated power for the trial, **Aveux** = naive mean in group x, **Aveuy** = naive mean in group y, and **AveN** = the average sample-size per group.

The key algorithms for SAS Macro 6.1 are specified as follows:

(1) Take inputs: **nSims**, **Model**, **nStags**, **Ns{i}**, **alpha{i}**, **beta {i}**, **ux**, **uy**, **NId**, **sigma**, **nAdj**, **DuHa**, **Nmax**, **N0.**

(2) Generate stagewise means **uxObs** and **uyObs** for the two groups (not individual patient response).

(3) Compute the test statistic **TS** for either MPI, MSP, or MPP.

(4) Check if TS crosses the stopping boundary.

If the boundary is crossed, update the power and the stopping probability **ESP{i}** and/or **FSP{i}**;

otherwise continue to the next stage of the trial.

(5) Loop back to step 2.

>>SAS Macro 6.1: N-Stage Adaptive Designs with Normal Endpoint>>

```
%Macro NStgAdpDsgNor(nSims=1000000, Model="MIP", nStgs=3,
        ux=0, uy=1, NId=0, sigma=2, nAdj="Y", DuHa=1,
        Nmax=200, N0=150);
DATA NStgAdpDsg; Set dInput;
KEEP power Aveux Aveuy AveN FSP1-FSP&nStgs
    ESP1-ESP&nStgs alpha1-alpha&nStgs beta1-beta&nStgs;
Array Ns{&nStgs}; Array alpha{&nStgs}; Array beta{&nStgs};
Array ESP{&nStgs}; Array FSP{&nStgs};
seedx=3637; seedy=1624; nStgs=&nStgs; sigma=&sigma;
power=0; AveN=0; Aveux=0; Aveuy=0; du=abs(&uy-&ux);
cumN=0; Do i=1 To nStgs-1; cumN=cumN+Ns{i}; End;
Do i=1 To nStgs; FSP{i}=0; ESP{i}=0; End;
```

```
Do iSim=1 to &nSims;
 ThisN=0; Thisux=0; Thisuy=0;
 TS=0; If &Model="MPP" Then TS=1;
 Do i=1 To nStgs;
     uxObs=Rannor(seedx)*sigma/Sqrt(Ns{i})+&ux;
     uyObs=Rannor(seedy)*sigma/Sqrt(Ns{i})+&uy;
     Thisux=Thisux+uxObs*Ns{i};
     Thisuy=Thisuy+uyObs*Ns{i};
     ThisN=ThisN+Ns{i};
     Z = (uyObs-uxObs+&NId)*Sqrt(Ns{i}/2)/sigma;
     pi=1-ProbNorm(Z);
     If &Model="MIP" Then TS=pi;
     If &Model="MSP" Then TS=TS+pi;
     If &Model="MPP" Then TS=TS*pi;
     If TS>beta{i} Then Do; FSP{i}=FSP{i}+1/&nSims;
        Goto Jump; End;
     Else If TS<=alpha{i} then do;
        Power=Power+1/&nSims; ESP{i}=ESP{i}+1/&nSims;
        Goto Jump; End;
     Else If i=1 & &Nadj="Y" Then Do;
       eRatio=&DuHa/(abs(uyObs-uxObs)+0.0000001);
       nFinal=min(&Nmax,max(&N0,eRatio*2*&N0));
       If nStgs>1 Then Ns{nStgs}= nFinal-cumN; End;
 End;
Jump:
Aveux=Aveux+Thisux/ThisN/&nSims;
Aveuy=Aveuy+Thisuy/ThisN/&nSims;
AveN=AveN+ThisN/&nSims;
End;
Output;
Run;
Proc Print; run;
%Mend NStgAdpDsgNor;
```

‹‹SAS‹‹

Example 6.1 Three-Stage Adaptive Design

In a phase-III asthma study with two dose groups (control and active), the primary efficacy endpoint is the percent change from baseline in FEV1. The estimated FEV1 improvement from baseline is 5% and 12% for the control and active groups, respectively, with a common standard deviation

of $\sigma = 22\%$.

We will discuss three-stage, group sequential designs with and without SSR using MSP. There are four simple steps to follow in order to design the trial with SAS Macro 6.1: (1) determine the stopping boundaries; (2) determine the power or sample-size; (3) perform sensitivity analysis; and (4) perform estimation.

Let's first illustrate the steps using the design without SSR.

(1) Determination of stopping boundary

Choose the number of analyses and the initial stagewise sample-size and stopping boundaries alpha{i} and beta{i}. Set the null hypothesis condition (e.g., ux = 0.05, uy = 0.05). Use the following SAS macro call to calculate the power under this null condition; then adjust the value of alpha{i} and beta{i} until the power = type-I error α.

```
>>SAS>>
Data dInput;
Array Ns{3} (100, 100, 100); Array alpha{3} (0.014,0.15,0.291);
Array beta{3} (1,1,1);
%NStgAdpDsgNor(Model="MSP", nStgs=3, ux=0.05, uy=0.05,
sigma=0.22, nAdj="N"); Run;
<<SAS<<
```

(2) Determination of the sample-size

Keep everything the same, but change the treatment effect to the alternative hypothesis. The following is the SAS macro call for the power calculation:

```
>>SAS>>
Data dInput;
Array Ns{3} (100, 100, 100); Array alpha{3} (0.014,0.15,0.291);
Array beta{3} (1,1,1);
%NStgAdpDsgNor(Model="MSP", nStgs=3,ux=0.05, uy=0.12,
sigma=0.22, nAdj="N"); Run;
<<SAS<<
```

If the power is different from the desired power, change the sample-size and redetermine the stopping boundaries and simulate the power again. The iteration process continues until the desired power is reached.

(3) Sensitivity analysis

Because the treatment difference and its variability are not exactly known, it is necessary to run the simulation under other critical conditions, which is referred to as sensitivity analysis or risk assessment.

The example of sensitivity analysis with the control mean $u_x = 0.05$ and the test mean $u_y = 0.11$ is given by the following SAS statement:

```
>>SAS>>
Data dInput;
Array Ns{3} (100, 100, 100); Array alpha{3} (0.014,0.15,0.291);
Array beta{3} (1,1,1);
%NStgAdpDsgNor(Model="MSP", nStgs=3,ux=0.05, uy=0.11,
sigma=0.22, nAdj="N"); Run;
<<SAS<<
```

The simulation results or the operating characteristics of the design are presented in Table 6.1.

Table 6.1: Operating Characteristics

Case	ESP1	ESP2	ESP3	FSP1	FSP2	FSP3	Power	$\bar{N}$
Ho	0.014	0.009	0.002	0.000	0.500	0.326	0.025	246
Ha	0.520	0.294	0.078	0.000	0.001	0.002	0.892	166
Hs	0.394	0.293	0.094	0.000	0.003	0.008	0.780	192

The stopping probabilities can be used to calculate the expected duration of the trial. In the current case, the conditional (on the efficacy claim) expected trial duration is given by

$$\bar{t}_e = \sum_{i=1}^{3} ESP\{i\}\, t_i,$$

where t_i is the time from the first-patient-in to the i^{th} interim analysis.

The conditional (on the futility claim) expected trial duration is given by

$$\bar{t}_f = \sum_{i=1}^{3} FSP\{i\}\, t_i.$$

The unconditional expected trial duration for the trial is given by

$$\bar{t} = \sum_{i=1}^{3} \left(ESP\{i\} + FSP\{i\}\right) t_i.$$

(4) Naive point estimations:

The average naive point estimate can be obtained using the SAS macro. Under the null hypothesis H_o : $u_x = u_y = 0.05$, the naive means are $\hat{u}_x = 0.0510$ and $\hat{u}_y = 0.0490$ for the control and test groups, respectively. We can see that the bias is negligible. Under the alternative H_a: $u_x = 0.05$ and $u_y = 0.12$. The naive mean estimates are $\hat{u}_x = 0.0462$ and $\hat{u}_y = 0.1240$. Under H_s: $u_x = 0.05$ and $u_y = 0.11$, the naive means are $\hat{u}_x = 0.0460$ and $\hat{u}_y = 0.1141$ for the two groups.

Similarly, for a three-stage design with SSR using MSP, the SAS macro calls for the design are given as follows.

(1) Determine the stopping boundaries

Using the following SAS macro call to determine the stopping boundaries:

```
>>SAS>>
Data dInput;
Array Ns{3} (100, 100, 100); Array alpha{3} (0.014,0.15,0.291);
Array beta{3} (1,1,1);
%NStgAdpDsgNor( Model="MSP", nStgs=3,ux=0.05, uy=0.05,
sigma=0.22, nAdj="Y", Nmax=500, N0=300);
<<SAS<<
```

By trial-error method, we can find that the stopping boundaries are virtually the same as those without SSR.

(2) Determine sample-size

Use the following SAS macro call to determine the sample-size required for the desired power:

```
>>SAS>>
Data dInput;
Array Ns{3} (100, 100, 100); Array alpha{3} (0.014,0.15,0.291);
Array beta{3} (1,1,1);
%NStgAdpDsgNor(Model="MSP", nStgs=3,ux=0.05, uy=0.12,
sigma=0.22, nAdj="Y", Nmax=500, N0=300);
<<SAS<<
```

(3) Sensitivity assessment

Use the following SAS call for the sensitivity analysis:

```
>>SAS>>
Data dInput;
Array Ns{3} (100, 100, 100); Array alpha{3} (0.014,0.15,0.291);
```

```
Array beta{3} (1,1,1);
%NStgAdpDsgNor(Model="MSP", nStgs=3,ux=0.05, uy=0.11,
sigma=0.22, nAdj="Y", Nmax=500, N0=300);
‹‹SAS‹‹
```

(4) Operating characteristics

The operating characteristics are summarized in Table 6.2.

Table 6.2: Operating Characteristics

Case	ESP1	ESP2	ESP3	FSP1	FSP2	FSP3	Power	$\bar{N}$
Ho	.014	.009	.002	.000	.500	.326	.025	342
Ha	.520	.294	.100	.000	.001	.000	.914	204
Hs	.394	.292	.137	.000	.003	.001	.823	254

Note that the operating characteristics are virtually the same under H_o, with or without SSR ($N_{max} = 500$). The sample adjustment is performed at the first interim analysis. Under H_s (ux = 5% vs. uy = 11%), the power increases by 4.3% (from 78% to 82.3%) with SSR compared to without SSR.

Naive point estimations:

The naive mean estimates are ($\hat{u}_x = 0.0521$, $\hat{u}_y = 0.0480$) under H_o, ($\hat{u}_x = 0.0455, \hat{u}_y = 0.1247$) under H_a; and ($\hat{u}_x = 0.0450$, $\hat{u}_y = 0.1150$) under H_s. These numbers indicate the magnitude of the potential bias caused by an adaptive design.

6.2.2 *Binary Endpoint*

SAS Macro 6.2 can be used for two-arm, N-stage adaptive designs with binary endpoints. The SAS variables are defined as follows: **nSims** = number of simulation runs, and **Model** = adaptive design method: "MIP," "MSP," or "MPP." **nStgs** = number of stages, **Px**, **Py** = response rates for groups x and y, respectively, **NId** = non-inferiority margin, and **sigma** = standard deviation. **nAdj** = "Y" to allow for sample-size adjustment; otherwise, **nAdj** = "N." **N0** = initial sample-size for the final stage, and **Nmax** = maximum sample-size allowed. **DuHa** = treatment difference; for power calculation, **DuHa** = **Py** - **Px**, and for sensitivity analysis, **DuHa** $\neq$ **Py** - **Px**. **Ns{i}** = the i^{th} stage sample-size (not cumulative one), **alpha{i}** = the efficacy stopping boundary at the i^{th} stage, and **beta{i}** = the futility stopping boundary at the i^{th} stage. **ESP{i}** = the efficacy stopping probability at the i^{th} stage, and **FSP{i}** = the futility stopping probability at the i^{th} stage. **power** = the simulated power for the trial,

AvePx = naive average response rate in group x, **AvePy** = naive average response rate in group y, and **AveN** = the average sample-size per group.

The key algorithms for SAS Macro 6.2 are specified as follows:

(1) Take inputs: **nSims**, **Model**, **nStags**, **Ns{i}**, **alpha{i}**, **beta {i}**, **px**, **py**, **NId**, **nAdj**, **DuHa**, **Nmax**, **N0.**

(2) Generate stagewise response rates **pxObs** and **pyObs** for the two groups.

(3) Compute the test statistic **TS** for MPI, MSP, or MPP.

(4) Check if TS crosses the stopping boundary.

If the boundary is crossed, update the power and the stopping probability **ESP{i}** or **FSP{i}**;

otherwise continue to the next stage of the trial.

(5) Loop back to step 2.

››SAS Macro 6.2: N-Stage Adaptive Designs with Binary Endpoint››

```
%Macro NStgAdpDsgBin(nSims=1000000, Model="MIP",nStgs=3,
 Px=0, Py=1, NId=0, nAdj="Y", DuHa=1,Nmax=200, N0=150);
DATA NStgAdpDsg; Set dInput;
KEEP power AvePx AvePy AveN FSP1-FSP&nStgs
     ESP1-ESP&nStgs alpha1-alpha&nStgs beta1-beta&nStgs;
Array Ns{&nStgs}; Array alpha{&nStgs}; Array beta{&nStgs};
Array ESP{&nStgs}; Array FSP{&nStgs};
seedx=3637; seedy=1624; nStgs=&nStgs; Px=&Px; Py=&Py;
power=0; AveN=0; AvePx=0; AvePy=0; cumN=0;
Do i=1 To nStgs-1; cumN=cumN+Ns{i}; End;
Do i=1 To nStgs; FSP{i}=0; ESP{i}=0; End;
Do iSim=1 to &nSims;
 ThisN=0; ThisPx=0; ThisPy=0;
 TS=0; If &Model="MPP" Then TS=1;
ThisN=0;
 Do i=1 To nStgs;
      PxObs=RanBin(seedx,Ns(i),Px)/Ns(i);
     PyObs=RanBin(seedy,Ns(i),Py)/Ns(i);
     ThisPx=ThisPx+PxObs*Ns(i);
     ThisPy=ThisPy+PyObs*Ns(i);
     ThisN=ThisN+Ns{i};
     sigma=((PxObs*(1-PxObs)+PyObs*(1-PyObs))/2)**0.5;
     Z = (PyObs-PxObs+&NId)*Sqrt(Ns{i}/2)/sigma;
     pi=1-ProbNorm(Z);
```

```
    If &Model="MIP" Then TS=pi;
    If &Model="MSP" Then TS=TS+pi;
    If &Model="MPP" Then TS=TS*pi;
    If TS>beta{i} Then Do; FSP{i}=FSP{i}+1/&nSims;
                    Goto Jump; End;
    Else If TS<=alpha{i} then do;
        Power=Power+1/&nSims; ESP{i}=ESP{i}+1/&nSims;
                    Goto Jump; End;
    Else If i=1 & &Nadj="Y" Then Do;
        eRatio=&DuHa/(abs(PyObs-PxObs)+0.0000001);
        nFinal=round(min(&Nmax,max(&N0,eRatio*2*&N0)));
        If nStgs>1 Then Ns{nStgs}= nFinal-cumN; End;
 End;
Jump:
AvePx=AvePx+ThisPx/ThisN/&nSims;
AvePy=AvePy+ThisPy/ThisN/&nSims;
AveN=AveN+ThisN/&nSims;
END;
output;
RUN;
proc print; run;
%Mend NStgAdpDsgBin;
```

‹‹SAS‹‹

Example 6.2 Four-Stage Adaptive Design

A phase III trial is to be designed for patients with acute ischemic stroke of recent onset. The composite endpoint (death and MI) is the primary endpoint and the event rate is 14% for the control group and 12% for the test group. The sample-size for a classic design is 5937 per group, which will provide 90% power at a one-sided alpha = 0.025.

For illustration purpose, we choose a four-stage design with and without SSR using MSP. Again, there are four simple steps to follow in order to design the trial with this SAS macro: (1) determine the stopping boundaries, (2) determine the power or sample-size, (3) perform sensitivity analysis, and (4) perform estimation.

Let's first discuss the design without SSR.

(1) Determination of stopping boundaries

Choose the number of analyses and the initial stagewise sample-size, and stopping boundaries **alpha{i}** and **beta{i}**. Set the null hypothesis condition (e.g., Px = 0.14, Py = 0.14) and repeat the simulation using the

following SAS macro call with different values of alpha{i} and beta{i} until the simulated power = type-I error α.

```
>>SAS>>
Data dInput;
Array Ns{4} (1500, 1500, 1500, 1500); Array beta{4} (1,1,1,1);
Array alpha{4} (0.002,0.0011,0.0003, 0.00011);
%NStgAdpDsgBin( Model="MPP", nStgs=4,Px=0.14, Py=0.14,
DuHa=0.02, Nmax=10000, N0=6000); Run;
<<SAS<<
```

(2) Determination of sample-size

Keep everything the same, but change the treatment effect to the alternative hypothesis; then submit the following SAS statement to obtain the power:

```
>>SAS>>
Data dInput;
Array Ns{4} (1500, 1500, 1500, 1500);  Array beta{4} (1,1,1,1);
Array alpha{4} (0.002, 0.0011, 0.0003, 0.00011);
%NStgAdpDsgBin(Model = "MPP", nStgs=4, Px=0.12, Py=0.14,
DuHa=0.02, Nmax=10000, N0=6000); Run;
<<SAS<<
```

(3) Operating characteristics

The operating characteristics are summarized in Table 6.3.

Table 6.3: Operating Characteristics without SSR

Case	ESP1	ESP2	ESP3	ESP4	Power	$\bar{\text{N}}$	$\bar{\text{u}}_x$	$\bar{\text{u}}_y$
Ho	.002	.007	.007	.009	.025	5959	.140	.140
Ha	.105	.333	.249	.168	.855	4155	.119	.141

(4) Naive point estimations:

The naive point estimates $\hat{u}_x$ and $\hat{u}_y$ are also presented in Table 6.3.

We now follow the same steps to simulate the designs with sample-size adjustment.

(1) Determination of stopping boundaries

Repeatedly submit the following SAS macro call with different values of alpha{i} and beta{i} to determine the stopping boundaries:

```
>>SAS>>
```

```
Data dInput;
Array Ns{4} (1500, 1500, 1500, 1500); Array beta{4} (1,1,1,1);
Array alpha{4} (0.002,0.0011,0.0003, 0.00011);
%NStgAdpDsgBin(Model="MPP", nStgs=4, Px=0.14, Py=0.14,
nAdj="Y",  DuHa=0.02, Nmax=10000, N0=6000); Run;
‹‹SAS‹‹
```

(2) Determine sample-size

Use the following SAS macro call to determine the sample-size required for the desired power:

```
››SAS››
Data dInput;
Array Ns{4} (1500, 1500, 1500, 1500); Array beta{4} (1,1,1,1);
Array alpha{4} (0.002,0.0011,0.0003, 0.00011);
%NStgAdpDsgBin( Model="MPP", nStgs=4, Px=0.12, Py=0.14,
nAdj="Y", DuHa=0.02, Nmax=10000, N0=6000); Run;
‹‹SAS‹‹
```

(3) Operating Characteristics

The operating characteristics are presented in Table 6.4. The power increases from 85.5% to 97.1% due to SSR.

Table 6.4: Operating Characteristics with SSR

Case	ESP1	ESP2	ESP3	ESP4	Power	$\bar{N}$	$\bar{u}_x$	$\bar{u}_y$
Ho	.002	.007	.007	.009	.025	9824	.140	.140
Ha	.105	0.333	.249	.284	.971	5374	.118	.142

6.2.3 *Survival Endpoint*

The SAS Macro 6.3 can be used for two-arm N-stage adaptive designs with binary, normal, or survival endpoints. The SAS variables are defined as follows: **nSims** = number of simulation runs, **nStgs** = number of stages, and **ux**, **uy** = means, response rates, or hazard rate for groups x and y, respectively, depending on the endpoint; **NId** = noninferiority margin; and **sigma** = standard deviation. **nAdj** = "Y" to allow for sample-size adjustment, otherwise, **nAdj** = "N", **N0** = initial cumulative sample-size for the final stage, and **Nmax** = maximum sample-size allowed. **DuHa** = treatment difference; for power calculation **DuHa** = **uy** - **ux**; for sensitivity analysis, **DuHa** $\neq$ **uy** - **ux**. **tAcr** = uniform accrual duration, **tStd** =

trial duration, **and Ns{i}** = the i^{th} stage sample-size (not cumulative one). **alpha{i}** = the efficacy stopping boundary at the i^{th} stage, **beta{i}** = the futility stopping boundary at the i^{th} stage. **ESP{i}** = the efficacy stopping probability at the i^{th} stage, and **FSP{i}** = the futility stopping probability at the i^{th} stage. **power** = the simulated power for the trial, **AvePx** = average response in group x, **AvePy** = average response in group y, and **AveN** = the average sample-size per group. **EP** = "normal", "binary", or "survival". **Model** is for the methods, which can be MIP, MSP, MPP, WZ, or UWZ. WZ is for the inverse-normal method with constant weights and UWZ is for the inverse-normal method with information time as the weights.

The key algorithms for SAS Macro 6.3 are specified as follows:

(1) Take inputs: **nSims**, **Model**, **nStags**, **Ns{i}**, **alpha{i}**, **beta {i}**, **ux**, **uy**, **NId**, **sigma**, **nAdj**, **DuHa**, **Nmax**, **N0, tAcr,** and **tStd.**

(2) Compute "standard deviation" sigma based on different endpoints.

(3) Generate stagewise response **uxObs**, and **uyObs** for the two groups.

(3) Compute the test statistic **TS** for MPI, MSP, MPP, WZ, and UWZ.

(4) Check if TS crosses the stopping boundary.

If the boundary is crossed, update the power and the stopping probability **ESP{i}** or **FSP{i};**

otherwise continue to the next stage of the trial.

(5) Loop back to step 2.

>>SAS Macro 6.3: N-Stage Adaptive Designs with Various Endpoint>>

```
%Macro TwoArmNStgAdpDsg(nSims=100000, nStgs=3,ux=0,
    uy=1, NId=0, EP="normal", Model="MSP", Nadj="N",
    DuHa=1, Nmax=300, N0=100, sigma=2, tAcr=10, tStd=24);
DATA NStgAdpDsg; Set dInput;
KEEP power Aveux Aveuy AveN FSP1-FSP&nStgs
    ESP1-ESP&nStgs alpha1-alpha&nStgs beta1-beta&nStgs;
Array Ns{&nStgs}; Array alpha{&nStgs}; Array beta{&nStgs};
Array ESP{&nStgs}; Array FSP{&nStgs}; Array Ws{&nStgs};
Array sumWs{&nStgs}; Array TSc{&nStgs};
seedx=3637; seedy=1624; Model=&Model; nStgs=&nStgs;
sigma=&sigma; power=0; AveN=0; Aveux=0; Aveuy=0;
cumN=0; Do i=1 To nStgs-1; cumN=cumN+Ns{i}; End;
Do k=1 To nStgs;
     sumWs{k}=0; Do i=1 To k;
                    sumWs{k}=sumWs{k}+Ws{i}**2; End;
```

```
    sumWs{k}=Sqrt(sumWs{k});
End;
* Calcate the standard deviation, sigma for different endpoints *;
 u=(&ux+&uy)/2;
 if &EP="normal" Then sigma=&sigma;
 if &EP="binary" Then sigma=(u*(1-u))**0.5;
 if &EP="survival" Then
    expud=exp(-u*&tStd);
    sigma=u*(1+expud*(1-exp(u*&tAcr))/(&tAcr*u))**(-0.5);
Do i=1 To nStgs; FSP{i}=0; ESP{i}=0; End;
Do iSim=1 to &nSims;
 ThisN=0; Thisux=0; Thisuy=0;
 Do i=1 To nStgs; TSc{i}=0; End;
 TS=0; If &Model="MPP" Then TS=1;
 Do i=1 To nStgs;
    uxObs=Rannor(seedx)*sigma/Sqrt(Ns{i})+&ux;
    uyObs=Rannor(seedy)*sigma/Sqrt(Ns{i})+&uy;
    Thisux=Thisux+uxObs*Ns{i};
    Thisuy=Thisuy+uyObs*Ns{i};
    ThisN=ThisN+Ns{i};
    TS0 = (uyObs-uxObs+&NId)*Sqrt(Ns{i}/2)/sigma;
If Model="MIP" Then TS=1-ProbNorm(TS0);
If Model="MSP" Then TS=TS+(1-ProbNorm(TS0));
If Model="MPP" Then TS=TS*(1-ProbNorm(TS0));
If Model="WZ" Then Do;
    Do k=i to nStgs;
        TSc{k}=TSc{k}+Ws{i}/sumWs{k}*TS0;
    End;
    TS=1-ProbNorm(TSc{i});
End;
If Model="UWZ" Then Do;
    TS0=((Thisuy-Thisux)/ThisN+&NId)*Sqrt(ThisN/2)/sigma;
    TS=1-ProbNorm(TS0);
End;
    If TS>beta{i} Then Do; FSP{i}=FSP{i}+1/&nSims;
                           Goto Jump; End;
    Else If TS<=alpha{i} then do;
        Power=Power+1/&nSims; ESP{i}=ESP{i}+1/&nSims;
                           Goto Jump; End;
    Else If i=1 & &Nadj="Y" Then Do;
        eRatio=&DuHa/(abs(uyObs-uxObs)+0.0000001);
```

```
            nFinal=min(&Nmax,max(&N0,eRatio*2*&N0));
            If nStgs>1 Then Ns{nStgs}= nFinal-cumN; End;
 End;
Jump:
Aveux=Aveux+Thisux/ThisN/&nSims;
Aveuy=Aveuy+Thisuy/ThisN/&nSims;
AveN=AveN+ThisN/&nSims;
End;
Output;
Run;
Proc Print; Run;
%Mend TwoArmNStgAdpDsg;
```
‹‹SAS‹‹

Example 6.3 Adaptive Design with Survival Endpoint

Consider a two-arm comparative oncology trial comparing a test drug to an active control with respect to the primary efficacy endpoint, time to disease progression (TTP). Based on data from previous studies, the median TTP is estimated to be 10 months (hazard rate = 0.0693) for the control group, and 13 months (hazard rate = 0.0533) for the test group. Assume that there is a uniform enrollment with an accrual period of 10 months and that the total study duration is expected to be 24 months. Sample-size calculation will be performed under the assumption of an exponential survival distribution.

To do the simulation, choose the number of analyses (K = 3) and the initial stagewise sample-size, and stopping boundaries alpha{i} and beta{i}. Define the null hypothesis condition (e.g., ux = 0.0693, Py = 0.0693). Similar to the steps in the previous two examples with Normal and binary endpoints, there are four simple steps to follow in order to design the trial with this SAS macro: (1) determine the stopping boundaries, (2) determine the power or sample-size, (3) perform sensitivity analysis, and (4) perform estimation. The corresponding SAS macro calls are presented as follows:

(1) Determination of stopping boundary

Use the following SAS macro call as an example to determine the stopping boundaries:

```
>>SAS>>
Data dInput;
Array Ns{3} (150, 150, 150); Array alpha{3} (0.002,0.01,0.02);
Array beta{3} (1,1,0.02); Array Ws{3} (1,1,1);
```

```
%TwoArmNStgAdpDsg(nStgs=3,ux=0.0693, uy=0.0693, EP="survival",
Model="WZ", tAcr=10, tStd=24); Run;
```
‹‹SAS‹‹

(2) Determination of sample-size

Keep everything the same, but change the treatment effect to the alternative hypothesis. Use the following SAS macro call to obtain the power:

```
>>SAS>>
Data dInput;
Array Ns{3} (150, 150, 150); Array alpha{3} (0.002,0.01,0.02);
Array beta{3} (1,1,0.02); Array Ws{3} (1,1,1);
%TwoArmNStgAdpDsg(nStgs=3,ux=0.0533, uy=0.0693, EP="survival",
Model="WZ", tAcr=10, tStd=24); Run;
```
‹‹SAS‹‹

(3) Sensitivity analysis

Use the following SAS macro call as an example for the sensitivity analysis:

```
>>SAS>>
Data dInput;
Array Ns{3} (150, 150, 150); Array alpha{3} (0.002,0.01,0.02);
Array beta{3} (1,1,0.02); Array Ws{3} (1,1,1);
%TwoArmNStgAdpDsg(nStgs=3,ux=0.0533, uy=0.066, EP="survival",
Model="WZ", tAcr=10, tStd=24); Run;
```
‹‹SAS‹‹

(4) Operating characteristics

The operating characteristics are summarized in Table 6.5.

Table 6.5: Operating Characteristics without Adjustment

Case	ESP1	ESP2	ESP3	Power	$\bar{N}$	$\bar{u}_x$	$\bar{u}_y$
Ho	.002	.009	.014	.025	448	.069	.069
Ha	.155	.474	.259	.888	332	.052	.070
Hs	.085	.345	.297	.727	373	.053	.067

Similarly, for the design with sample-size adjustment, the stopping boundaries, the initial sample-size determination and the sensitivity analysis can be carried out using the SAS macro calls as described below.

(1) Determination of Stopping Boundaries

Use the following SAS macro call as an example to determine the stopping boundaries:

```
>>SAS>>
Data dInput;
Array Ns{3} (150, 150, 150); Array alpha{3} (0.002,0.0075,0.02);
Array beta{3} (1,1,0.02); Array Ws{3} (1,1,1);
%TwoArmNStgAdpDsg(nStgs=3,ux=0.0693, uy=0.0693, EP="survival",
Model="UWZ", Nadj="Y", DuHa=0.016, Nmax=600, N0=450, tAcr=10,
tStd=24);
<<SAS<<
```

(2) Determination of sample-size

Keep everything the same, but change the treatment effect to the alternative hypothesis. Use the following SAS macro call to obtain the power:

```
>>SAS>>
Data dInput;
Array Ns{3} (150, 150, 150); Array alpha{3} (0.002,0.0075,0.02);
Array beta{3} (1,1,0.02); Array Ws{3} (1,1,1);
%TwoArmNStgAdpDsg(nStgs=3,ux=0.0533, uy=0.0693, EP="survival",
Model="UWZ", Nadj="Y", DuHa=0.016,  Nmax=600, N0=450, tAcr=10,
tStd=24);
<<SAS<<
```

(3) Sensitivity analysis

Use the following SAS macro call as an example for the sensitivity analysis:

```
>>SAS>>
Data dInput;
Array Ns{3} (150, 150, 150); Array alpha{3} (0.002,0.0075,0.02);
Array beta{3} (1,1,0.02); Array Ws{3} (1,1,1);
%TwoArmNStgAdpDsg(nStgs=3,ux=0.0533, uy=0.066, EP="survival",
Model="UWZ", Nadj="Y", DuHa=0.016, Nmax=600, N0=450, tAcr=10,
tStd=24);
<<SAS<<
```

(4) Operating characteristics

The operating characteristics are summarized in Table 6.6.

Table 6.6: Operating Characteristics with SSR

Case	ESP1	ESP2	ESP3	Power	$\bar{N}$	$\bar{u}_x$	$\bar{u}_y$
Ho	.002	.007	.016	.025	597	.069	.069
Ha	.155	.435	.366	.956	400	.052	.070
Hs	.085	.306	.449	.840	470	.052	.067

6.3 Error-Spending Approach

The error-spending approach requires prespecification of the error-spending function (Chapter 3). The advantage of this approach is that it allows for changes in the number and timing of analyses. To use the error-spending approach, simulations are performed with different stopping boundaries α_k and β_k until the efficacy stopping probabilities are equal to the corresponding prespecified value of the incremental error-spending function $\pi_k = \pi^*(k) - \pi^*(k-1)$. The commonly used error-spending functions are (1) O'Brien-Fleming: $\pi^*(I) = 2\left\{1 - \Phi(\frac{z_{1-\alpha/2}}{\sqrt{I}})\right\}$, (2) Pocock: $\pi^*(I) = \alpha \log[1 + (e-1)I]$, and (3) power family: $\pi^*(I) = \alpha I^\gamma, \gamma > 0$, where information time $I_k = \frac{n_k}{N}$ and $I_0 = 0$.

6.4 Summary

In this chapter we have demonstrated how to implement adaptive design with more than two stages using simulations, and we provided the SAS macros. We have shown you step by step how to use these macros to conduct trial designs. In determining the appropriate adaptive design, it is important to conduct sensitivity analyses and compare operating characteristics among different designs. In contrast to the simulation method, we will introduce you to the recursive adaptive design methods in Chapter 8, where you will find many closed forms for K-stage adaptive designs. However, before that we will discuss another interesting method used mainly for two-stage designs: the conditional error function method.

Problem

6.1 SAS Macro 6.1 is based on the randomly degenerated mean responses for individual groups. Please modify the Macro such that it is based on the randomly generated individual patient response, instead of mean responses; then compare the results from the two different approaches for small sample-size trials.

Chapter 7

Conditional Error Function Method

In this Chapter, we are going to discuss the so-called conditional error function method (CEFM), used mainly for two-stage designs. Researchers of this method include Proschan and Hunsberger (1995), Liu and Chi (2001), Müller and Schäfer (2001), and Denne (2001), among others.

7.1 Proschan-Hunsberger Method

Proschan and Hunsberger (1995) proposed a conditional error function method for two-stage design. Here we modify the Proschan-Hunsberger method slightly to fit different types of endpoints by using inverse-normal transformation $z_k = \Phi^{-1}(1-p_k)$ and $p_k = 1-\Phi(z_k)$, where p_k is the stagewise p-value based on a subsample from stage k.

Let the test statistics for the first stage (sample-size n_1) and second stage (sample-size n_2) be

$$T_1 = p_1 \tag{7.1}$$

and

$$T_2 = 1-\Phi\left(w_1\Phi^{-1}(1-p_1)+w_2\,\Phi^{-1}(1-p_2)\right), \tag{7.2}$$

respectively.

The stopping rules are given by

$$\begin{cases} \text{If } T_k \le \alpha_k,\ (k=1,2),\ \text{stop and reject } H_o; \\ \text{If } T_k > \beta_k,\ (k=1,2),\ \text{stop and accept } H_o \\ \text{Otherwise,} \qquad\qquad \text{Continue,} \end{cases} \tag{7.3}$$

where $\alpha_2 = \beta_2$.

Assume that T_1 has the standard normal distribution under the null hypothesis. Let $A(p_1)$ be the conditional probability of making type-I error

at the second stage given $T_1 = p_1$. Notice that p_1 is uniformly distributed over [0,1]; a level α test requires:

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\beta_1} A\left(p_1\right) dp_1, \tag{7.4}$$

where $A\left(p_1\right)$ is called the conditional error function on a p-scale, which is similar to the conditional error function on a z-scale given by Proschan and Hunsberger (1995). The conditional error function can be any nondecreasing function $0 \leq A\left(p_1\right) \leq 1$ as far as type-I error is concerned.

Proschan and Hunsberger (1995) suggested the circular conditional error function:

$$A(p_1) = \quad 1 - \Phi(\sqrt{[\Phi^{-1}\left(1-\beta_1\right)]^2 - \left[\Phi^{-1}\left(1-p_1\right)\right]^2}), \quad \alpha_1 < p_1 \leq \beta_1. \tag{7.5}$$

Let $\beta_1 = \alpha_2$. For a given α and a predetermined α_1, β_1 can be calculated numerically by substituting (7.5) into (7.4). For example, with a one-sided $\alpha = 0.025$, and $\alpha_1 = 0.0147$, β_1 will be 0.174.

The stopping boundaries are derived for a classic group sequential design, i.e., no other adaptations like sample-size re-estimation. With sample-size modification, the stopping boundaries are still valid if the weight w_k (7.2) is prefixed. However, if the sample-size at the second stage is dependent on the data from the first stage and the weights in (7.2) are not a constant, e.g., $w_i = \sqrt{n_i/\left(n_1 + n_2\right)}$, $(i = 1, 2)$, the stopping boundaries determined using the above method are not valid anymore. In such cases, Proschan and Hunsberger (1995) suggested modifying α_2 but leaving the conditional error function $A(p_1)$ unchanged, and consequently the test is still a level α test.

To determine α_2, let's first derive the conditional power and conditional error function.

Let $cP_\delta(n_2, \alpha_2|p_1)$ be the conditional power $\Pr\left(T_2 \leq \alpha_2|p_1, \delta\right)$, where δ is the effect size or treatment. Assuming a large sample-size and known variance σ^2, we can obtain conditional power (see Section 7.4):

$$cP_\delta(n_2, \alpha_2|p_1) = 1 - \Phi\left[\frac{\Phi^{-1}\left(1-\alpha_2\right) - w_1\Phi^{-1}\left(1-p_1\right)}{w_2} - \frac{\delta}{\sigma}\sqrt{\frac{n_2}{2}}\right], \tag{7.6}$$

where $w_1^2 + w_2^2 = 1$. For a given α_1 and β_1, the conditional type-I error can be obtained by letting $\delta = 0$ in (7.6):

$$A(p_1) = 1 - \Phi\left[\frac{\Phi^{-1}(1-\alpha_2) - w_2\Phi^{-1}(1-p_1)}{w_1}\right]. \tag{7.7}$$

Because $A(p_1)$ is the same with or without SSR, it can be obtained through the procedure described for no SSR; then solve (7.7) for α_2:

$$\alpha_2 = \frac{\sqrt{n_1}\Phi^{-1}(1-p_1) + \sqrt{n_2}\Phi^{-1}(1-A(p_1))}{\sqrt{n_1+n_2}}. \tag{7.8}$$

Note that we have used the equation $w_i = \sqrt{n_i/(n_1+n_2)}$. From (7.6) and (7.7), we can obtain the conditional power

$$cP_\delta(n_2, z_c|p_1) = 1 - \Phi\left(\Phi^{-1}(1-A(p_1)) - \frac{\delta}{\sigma}\sqrt{\frac{n_2}{2}}\right). \tag{7.9}$$

To achieve a target conditional power cP, the sample-size required can be obtained by solving (7.9) for n_2:

$$n_2 = \frac{2\sigma^2}{\delta^2}\left[\Phi^{-1}(1-A(p_1)) - \Phi^{-1}(1-cP)\right]^2. \tag{7.10}$$

Note that for constant conditional error, $A(p_1) = c$, where c is a constant, (7.5) leads to $\alpha = \alpha_1 + c(\beta_1 - \alpha_1)$. Therefore the constant conditional error approach is equivalent to MIP.

Example 7.1 Adaptive Design for Coronary Heart Disease Trial

Suppose we are interested in a clinical trial in patients with coronary heart disease (CHD) that compares a cholesterol-reducing drug to a placebo with respect to angiographic changes from baseline to end of study (Proschan and Hunsberger, 1995, p.77). The coronary arteries are first divided into segments; for each segment the difference in minimum lumen diameter from baseline to end of study is computed, and the average difference over all segments of a patient is the outcome measure. It is not known what constitutes a minimum clinically relevant change, but another similar study showed an effect size of about one third of the observed standard deviation. The sample-size required for 90% power to detect a similar effect size is about 190 patients per group. It has been predetermined that the circular conditional error function will be used with one interim analysis based on evaluations of 95 patients in each arm. If $z_1 > 2.27$

($p_1 < 0.0116$), the trial will be stopped for efficacy. If $z_1 < 0$ ($p_1 > 0.5$) at the interim analysis, the trial will be stopped for futility. If $0 \leq z_1 < 2.27$ ($0.0116 < p_1 \leq 0.5$), proceed to the second stage. Suppose after the interim analysis, the z-score $z_1 = 1.5$ ($p_1 = 0.0668$). The corresponding effect size (δ/σ) is about 0.218. The conditional error is $\mathrm{A}(0.0668) = 0.0436$ which is obtained from (7.7) with $w_1 = w_2 = \sqrt{0.5}$ and $\alpha_2 = 0.0116$. To have 80% conditional power under the empirically estimated treatment effect, the newly estimated sample-size for the second stage is 274/group from (7.10).

7.2 Denne Method

Denne (Denne, 2001) has developed a new procedure for SSR at an interim analysis. Instead of keeping the conditional error ($\int_{\alpha_1}^{\beta_1} A(p_1)\, dp_1$) constant when making an adaptation, the Denne method ensures that the conditional error function $A(p_1)$ remains unchanged.

Let w_{01}, w_{02}, and α_{02} be weights and the final stopping boundary before sample-size modification; let w_1, w_2, and α_2 be weights and the final stopping boundary after sample-size modification. To control overall α, the stopping boundary α_2 is adjusted such that $A(p_1)$ is unchanged:

$$\frac{\Phi^{-1}(1-\alpha_{02}) - w_{02}\Phi^{-1}(1-p_1)}{w_{01}} = \frac{\Phi^{-1}(1-\alpha_2) - w_2\Phi^{-1}(1-p_1)}{w_1}, \tag{7.11}$$

where $w_{0i} = \sqrt{n_{0i}/(n_{01}+n_{02})}$ and $\tilde{w}_i = \sqrt{n_i/(n_1+n_2)}$; n_{0i} and n_i are the subsample size at the i^{th} stage before and after sample-size modification ($n_{01} = n_1$).

Equation (7.11) can be solved for α_2:

$$\alpha_2 = 1 - \Phi\left[\frac{w_1}{w_{01}}\left\{\Phi^{-1}(1-\alpha_{02}) - w_{02}\Phi^{-1}(1-p_1)\right\} + w_2\Phi^{-1}(1-p_1)\right]. \tag{7.12}$$

Denne also stated that we can first modify the sample-size based on the estimated variance at the first stage before unblinding the data, without modifying the stopping boundary. In such cases, the subsample size n_2 is the sample-size after modification based on blinded data.

Note that Denne's method is originally based on a z-scale; hence, the stopping boundary is given by

$$c_2 = \frac{w_1}{w_{01}} \{c_{02} - w_{02} z_1\} + w_2 z_1, \tag{7.13}$$

where c_{02} and c_2 are the original and modified final stopping boundaries.

7.3 Müller-Schäfer Method

The procedure developed by Müller and Schäfer (2001) is based on calculating conditional rejection error probabilities for classical group sequential designs with any number of stages. The conditional rejection error probability is the probability that the null-hypothesis will be rejected at a future stage of the design, given the value of the test statistic at an interim analysis, if the null hypothesis is true. Thereby, every choice of n-stage group sequential boundaries in the usual model of a Brownian motion process implicitly defines a conditional error function from which the type-I error risk for the rest of the trial after the interim analysis can be obtained (Müller and Schäfer, 2004). The Müller-Schäfer procedure can be viewed as a special case of the general concept developed by Müller and Schäfer (2004) applied to conventional group sequential designs in the Brownian motion model at the pre-determined time points of the interim analyses. Müller and Schäfer (2001) showed, by combining the method with the product of p-values and the method with Brownian motion, how one can make any data dependent change in an on-going adaptive trial and still preserve the overall type-I error. To achieve this, all one need do is preserve the conditional type-I error of the remaining portion of the trial. The conditional error usually can be calculated in real time for a given observed treatment difference.

We will discuss the trial examples of Müller-Schäfer in Chapter 8 as special cases of recursive two-stage adaptive designs.

In the next section, we will compare different methods based on their conditional error and conditional power functions.

7.4 Comparison of Conditional Power

Conditional power is a very useful operating characteristic for adaptive designs. It can be used for interim decision-making, and to make comparisons among different designs and statistical methods. Because the stopping boundaries for the most existing methods are either based on a z-scale or a p-scale, for the purpose of comparisons later, we will convert them all to p-

scale using the following simple transformation: $p_k = 1 - \Phi(z_k)$. Inversely, we have $z_k = \Phi^{-1}(1 - p_k)$, where z_k is the z-score from the subsample, which has an asymptotically normal distribution with $N(\delta/se\left(\hat{\delta}_2\right), 1)$ under the alternative hypothesis, where $\hat{\delta}_2$ is an estimation of treatment difference in the second stage and $se\left(\hat{\delta}_2\right) = \sqrt{\frac{2\hat{\sigma}^2}{n_2}} \approx \sqrt{\frac{2\sigma^2}{n_2}}$. To derive the conditional power, we express the criterion for rejecting H_o as

$$z_2 \geq B(\alpha_2, p_1). \tag{7.14}$$

From (7.14), we can immediately obtain the conditional probability at the second stage:

$$cP_\delta(p_1) = 1 - \Phi\left(B(\alpha_2, p_1) - \frac{\delta}{\sigma}\sqrt{\frac{n_2}{2}}\right), \alpha_1 < p_1 \leq \beta_1. \tag{7.15}$$

For Fisher's combination method, the rejection criterion for the second stage is $p_1 p_2 \leq \alpha_2$, i.e., $z_2 \geq \Phi^{-1}\left(1 - \frac{\alpha_2}{p_1}\right)$. Therefore, $B(\alpha_2, p_1) = \Phi^{-1}\left(1 - \frac{\alpha_2}{p_1}\right)$. Similarly, for the method based on the sum of stagewise p-values, the rejection criterion for the second stage is $p_1 + p_2 \leq \alpha_2$, i.e., $z_2 = B(\alpha_2, p_1) = \Phi^{-1}(1 - \max(0, \alpha_2 - p_1))$. For the inverse-normal method, the rejection criterion for the second stage is $w_1 z_1 + w_2 z_2 \geq \Phi^{-1}(1 - \alpha_2)$, i.e., $z_2 \geq \frac{\Phi^{-1}(1-\alpha_2) - w_1 \Phi^{-1}(1-p_1)}{w_2}$, where w_1 and w_2 are prefixed weight satisfying the condition: $w_1^2 + w_2^2 = 1$. Note that the group sequential design and the CHW method (Cui, Hung, and Wang, 1999) are special cases of the inverse-normal method.

The conditional error can be obtained by setting $\delta = 0$ in (7.15):

$$A(p_1) = 1 - \Phi(B(\alpha_2, p_1)), \alpha_1 < p_1 \leq \beta_1, \tag{7.16}$$

where the functions $B(\alpha_2, z_1)$ are summarized in Table 7.1 for different design methods.

Substituting (7.16) into (7.4), we can obtain the following formulation for determining the stopping boundaries for various designs:

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\beta_1} 1 - \Phi(B(\alpha_2, p_1))\, dp_1. \tag{7.17}$$

If the trial continues, i.e., $\alpha_1 < p_1 \leq \beta_1$, for given conditional power cP, we can solve (7.15) for the adjusted sample-size for the second stage:

$$n_2 = \frac{2\sigma^2}{\delta^2}\left(B(\alpha_2, z_1) - \Phi^{-1}(1 - cP)\right)^2. \tag{7.18}$$

Table 7.1: Function $B(\alpha_2, p_1)$ for Conditional Power

Design	$B(\alpha_2, p_1)$
Sequential Design	$\frac{z_{1-\alpha_2}-\sqrt{f_1^o}\Phi^{-1}(1-p_1)}{\sqrt{f_2^o}}$
MIP (Chang, 2006)	$\Phi^{-1}(1-\alpha_2)$
MSP (Chang, 2006)	$\Phi^{-1}(1-\max(0,\alpha_2-p_1))$
MPP, Bauer, and Kohne (1994)	$\Phi^{-1}\left(1-\frac{\alpha_2}{p_1}\right)$
Proschan-Hunsberger (1995)	$\sqrt{z_{1-\beta_1}^2 - z_{1-p_1}^2}$
Fisher-Shen Method (1998, 1999)	$\frac{\Phi^{-1}(1-\alpha)-w_1\Phi^{-1}(1-p_1)}{w_2}$
Lehmacher-Wassmer (1999)	$\frac{\Phi^{-1}(1-\alpha_2)-w_1\Phi^{-1}(1-p_1)}{w_2}$
Denne (2001)	$\frac{\Phi^{-1}(1-\alpha_2)-w_1\Phi^{-1}(1-p_1)}{w_2}$
Cui-Hung-Wang (1999)	$\frac{z_{1-\alpha_2}-\sqrt{f_1^o}\Phi^{-1}(1-p_1)}{\sqrt{f_2^o}}$

$\delta = (\mu_2 - \mu_1)/\sigma$, $n_2 = n$ per group at stage 2. Assume known σ.
f_i^o = initial sample-size fraction and $w_1^2 + w_2^2 = 1$.

For convenience, the conditional power (7.15) is implemented in SAS Macro 7.1. The SAS variables are defined as follows: the endpoint, **EP** = "normal" or "binary"; **ux** and **uy** = the responses (means or proportions) for the two groups, respectively, **sigma** = standard deviation for the normal endpoint; **Model** ="MIP", "MSP", "MPP", or "LW" for the four methods in Table 7.1; **alpha2** = the efficacy boundary at the second stage; **cPower** = the conditional power; **p1** = the stagewise p-value at the first stage; **w1** and **w2** = weights for Lehmacher-Wassmer method; and **n2** = sample-size per group for the second stage.

››SAS Macro 7.1: Conditional Power››

```
%Macro ConPower(EP="normal", Model="MSP", alpha2=0.205,
 ux=0.2, uy=0.4, sigma=1, n2=100, p1=0.8, w1=1, w2=1);
** cPower=Two stage conditional power. eSize=delta/sigma;
data cPower;
 a2=&alpha2; Model=&Model;
 u=(&ux+&uy)/2;
 w1=&w1/sqrt(&w1**2+&w2**2);
 w2=&w2/sqrt(&w1**2+&w2**2);
 If &EP="normal" Then sigma=&sigma;
 If &EP="binary" Then sigma=(u*(1-u))**0.5;
 eSize=(&uy-&ux)/sigma;
If Model="MIP" Then BFun=Probit(1-a2);
```

```
If Model="MSP" Then BFun=Probit(1-max(0.0000001,a2-&p1));
If Model="MPP" Then BFun=Probit(1-a2/&p1);
If Model="LW" Then BFun=(Probit(1-a2)- w1*Probit(1-&p1))/w2;
cPower=1-ProbNorm(BFun-eSize*sqrt(&n2/2));
Run;
Proc Print data=cPower; Run;
%Mend ConPower;
```
‹‹SAS‹‹

An example of determining the sample-size based on conditional power using SAS Macro 7.1 is given below:

‹‹SAS‹‹
```
%ConPower(EP="binary", Model="MSP", alpha2=0.2050,
ux=0.2, uy=0.4, n2=100, p1=0.1);
%ConPower(EP="binary", Model="MIP", alpha2=0.0201,
ux=0.2, uy=0.4, n2=100, p1=0.1);
%ConPower(EP="binary", Model="MPP", alpha2=0.0043,
ux=0.2, uy=0.4, n2=100, p1=0.1);
%ConPower(EP="binary", Model="LW", alpha2=0.0226,
ux=0.2, uy=0.4, n2=100, p1=0.1, w1=1, w2=1);
%ConPower(EP="normal", Model="MSP", alpha2=0.2050,
ux=0.2, uy=0.4, sigma=1, n2=200, p1=0.1);
%ConPower(EP="normal", Model="MIP", alpha2=0.0201,
ux=0.2, uy=0.4, sigma=1, n2=200, p1=0.1);
%ConPower(EP="normal", Model="MPP", alpha2=0.0043,
ux=0.2, uy=0.4, sigma=1, n2=200, p1=0.1);
%ConPower(EP="normal", Model="LW", alpha2=0.0226,
ux=0.2, uy=0.4, sigma=1, n2=200, p1=0.1, w1=1, w2=1);
```
‹‹SAS‹‹

The results with a non-binding futility boundary are presented in Table 7.2. We can see that MSP produces the highest power. It is important to know that in adaptive design, the conditional power is more important than unconditional power.

Table 7.2: Comparisons of Conditional Powers cP

	Endpoint	MSP	MIP	MPP	LM
α_2		0.2050	0.0210	0.0043	0.0226
cP (n = 100)	binary	0.967	0.850	0.915	0.938
cP (n = 200)	normal	0.772	0.479	0.611	0.673

Note: $\alpha_1 = 0.005$, $\beta_1 = 0.25$,and $p_1 = 0.01$.

The sample-size required at the second stage based on the conditional power (7.15) is also implemented in SAS Macro 7.2. The SAS variables are defined as follows: **nAdjModel** ="MIP", "MSP", "MPP", or "LW" for the four methods in Table 7.1: **alpha2** = the efficacy boundary at the second stage; **eSize** = standard effect size; **cPower** = the conditional power; **p1** = the stagewise p-value at the first stage; **w1** and **w2** = weights for Lehmacher-Wassmer method; and **n2New** = new sample-size required for the second stage to achieve the desired conditional power.

>>SAS Macro 7.2: Sample-Size Based on Conditional Power>>

```
%Macro nByCPower(nAdjModel, alpha2, eSize,
      cPower, p1, w1, w2, n2New);
 a2=&alpha2;
 If &nAdjModel="MIP" Then BFun=Probit(1-a2);
 If &nAdjModel="MSP" Then BFun=Probit(1-max(0.0000001,a2-&p1));
 If &nAdjModel="MPP" Then BFun=Probit(1-a2/&p1);
 If &nAdjModel="LW" Then
     BFun=(Probit(1-a2)- &w1*Probit(1-&p1))/&w2;
 &n2New=2*((BFun-Probit(1-&cPower))/&eSize)**2; *n per group;
 %Mend nByCPower;
```

<<SAS<<

An example of determining the sample-size based on conditional power using SAS Macro 7.2 is given below:

>>SAS>>

```
Data cPow; keep n2New;
%nByCPower("MSP", 0.1840, 0.21, 0.8, 0.0311, 0.707, 0.707, n2New);
Run;
Proc Print; Run;
```

<<SAS<<

Based on the SAS output, the new sample-size required for the second stage is 158 per group.

Table 7.3: Conditional Error Functions

Method	Test Statistic at Stage 2 $f(p_1, p_2)$	Conditional Error Function $p_2(p_1, \alpha_2)$	Intersect α_0
MSP	p_1+p_2	$p_2= \alpha_2-p_1$	$p_1= \alpha_2$
MLP	$w_1p_1+w_2p_2$	$p_2=\frac{1}{w_2}(\alpha_2-w_1p_1)$	$p_1=\frac{\alpha_2}{w_1}$
MPP	p_1p_2	$p_2=\frac{\alpha_2}{p_1}$	$p_1= +\infty$
MCP	$\sqrt{p_1^2+p_2^2}$	$p_2=\sqrt{\alpha_2^2-p_1^2}$	$p_1= \alpha_2$
MNP	$\sqrt{(p_1-\beta_1)^2+(p_2-\alpha_2)^2}$; $\beta_1= \alpha_1+a_2$	$p_2= \alpha_2 +\sqrt{\alpha_2-(p_1-\beta_1)^2}$	$p_1= \beta_1$

Table 7.4: Type-I Error Rate at Stage 2

Method	Type-I Error Rate at Stage 2: π_2
MSP	$\alpha_2 \min(\alpha_2, \beta_1) -\alpha_1\alpha_2+\frac{1}{2}\alpha_1^2-\frac{1}{2}\min^2(\alpha_2, \beta_1)$
MLP	$\frac{1}{w_2}[\frac{1}{2}w_1\alpha_1^2-\alpha_1\alpha_2+\alpha_2 \min\left(\frac{1}{w_1}\alpha_2, \beta_1\right)-\frac{1}{2}w_1\min^2\left(\frac{1}{w_1}\alpha_2, \beta_1\right)]$
MPP	$\alpha_2 \ln\frac{\beta_1}{\alpha_1}$
MCP	$\frac{\min(\beta_1,\alpha_2)}{2}\sqrt{\alpha_2^2-\min^2(\beta_1,\alpha_2)}-\frac{\alpha_1}{2}\sqrt{\alpha_2^2-\alpha_1^2}$ $+\frac{\alpha_2^2}{2}\left(\sin^{-1}\frac{\min(\beta_1,\alpha_2)}{\alpha_2}-\sin^{-1}\frac{a_1}{\alpha_2}\right)$
MNP	$\left(1-\frac{\pi}{4}\right)\alpha_2^2$. where $\beta_1 = \alpha_1 + a_2$

Table 7.5: Stopping Boundaries without Futility Binding

Test Statistic at Stage 2 $f(p_1, p_2)$	Stopping Boundary Requirement (no futility binding)
p_1+p_2	$\alpha_1+\frac{1}{2}(\alpha_2-\alpha_1)^2= \alpha$
$w_1p_1+w_2p_2$	$\alpha_1+\frac{1}{2w_1w_2}[\alpha_2-w_1\alpha_1]^2= \alpha$
p_1p_2	$\alpha_1+\alpha_2\ln\frac{1}{\alpha_1}= \alpha$
$\sqrt{p_1^2+p_2^2}$	$\alpha_1+\frac{\alpha_2^2}{2}\left(\frac{\pi}{2}-\sin^{-1}\frac{a_1}{\alpha_2}\right)-\frac{\alpha_1}{2}\sqrt{\alpha_2^2-\alpha_1^2}= \alpha$
$\sqrt{(p_1-\beta_1)^2+(p_2-\alpha_2)^2}$	$\alpha_1+\left(1-\frac{\pi}{4}\right)\alpha_2^2= \alpha; \beta_1= \alpha_1+\alpha_2$ (*binding*)

We can easily develop more closed forms of stopping boundaries for two-stage adaptive designs. Let's denote the test statistic by $f(p_1, p_2)$ for the second stage and the rejection rule by $f(p_1, p_2) \leq \alpha_2$. From the equation $f(p_1, p_2) = \alpha_2$, we can solve for p_2, i.e., $p_2 = p_2(p_1, \alpha_2)$. Under the null condition, $p_2 = p_2(p_1, \alpha_2)$ is the conditional error function. Further more, the Type-I error rate at stage 2 is given by

$$\pi_2 = \int_{\alpha_1}^{\min(\beta_1,\alpha_0)} p_2(p_1, \alpha_2)\, dp_1,$$

where α_0 is the intersect of the stopping boundary as indicated in Figure 7.1. Therefore, a α-level test requires

$$\alpha = \pi_1 + \pi_2 = \alpha_1 + \int_{\alpha_1}^{\min(\beta_1, \alpha_0)} p_2\left(p_1, \alpha_2\right) dp_1.$$

The efficacy boundary α_2 is usually a constant, but it can be a function of p_1. The stopping boundaries without futility binding can be obtained by assigning $\beta_1 = 1$. Tables 7.3 through 7.5 summarize the results for some test statistics.

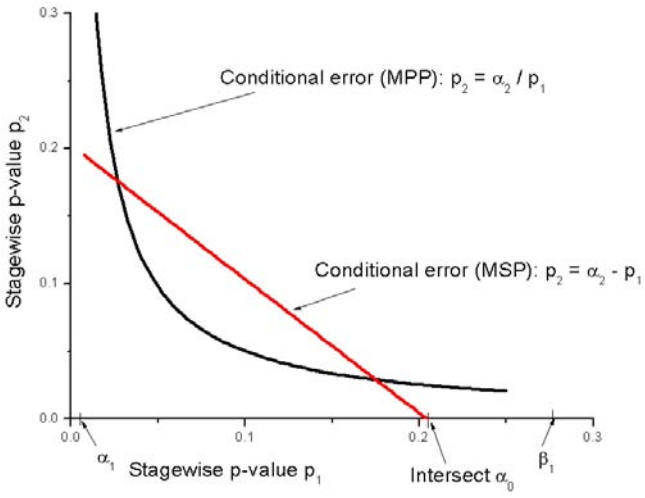


Figure 7.1: Conditional Error Functions

7.5 Adaptive Futility Design

7.5.1 *Utilization of an Early Futility Boundary*

In a trial with an early futility boundary, theoretically the final alpha for claiming statistic significance should increase. This may cause concern on the part of regulatory bodies, as the current practice is that the futility boundaries are not strictly followed by the sponsors, and therefore, the futility boundary should not be used to determine the later stopping boundaries. However, this is not very reasonable in the sense that we punish someone for others not following the rules. A more reasonable approach is that if the futility boundaries were actually followed, then the stopping boundaries

with futility boundaries considered can be used. If the futility boundaries were not followed, i.e., the trial was continued even though it had crossed the futilities boundaries, then the early futility boundary cannot be considered in constructing the later stopping boundaries. For example, in a two-stage trial with MSP, the stopping boundaries in a two-stage design are $\alpha_1 = 0.0025$, $\beta_1 = 0.15$, $\alpha_2 = 0.2288$. Suppose the stagewise p-value from the first stage was $p_1 = 0.1$, and the trial was continued to the second stage. Therefore, the futility boundary has been followed, and the final stopping boundary should be $\alpha_2 = 0.2288$. On the other hand, if $p_1 = 0.2$, and the trial was continued, the futility boundary is violated. Therefore, the final stopping boundary should be $\alpha_2 = 0.2146$ (corresponding to $\alpha_1 = 0.0025$, $\beta_1 = 1$), instead of 0.2288.

Note that MIP uses the early stopping boundary to construct the later stopping boundaries; this is different from traditional separate trials.

7.5.2 *Design with a Futility Index*

During any interim analysis, we should always perform futility checking and ask the question: Is it better to start a new trial than to continue the current one? When the conditional power is less than the power of the new trial with a sample-size equal to the adjusted sample-size for the current trial, it is better to start a new trial, theoretically. In other words, if the conditional power $cP_{\delta_1}(n_2)$ is less than the unconditional power $p_{\delta_1}(n_2)$, where n_2 is the sample-size for the second stage, then statistically, it is better to start a new trial. Alternatively, we can use a predetermined futility boundary to prevent the trial from continuing to the second stage when the conditional power is lower or the futility index is high.

7.6 Summary

Conditional error function methods (CEFM) allow for a broad selection of conditional error functions $A(p_1)$, as long as they are monotonic in the stagewise p-value p_1 and bound by [0,1]. The selection of different $A(p_1)$ implies using different weights for the data from the two stages. The CEFM requires keeping the conditional error unchanged when we make adaptations to the trial. The Muller-Schafer and Denne methods are special methods, i.e., they keep the conditional function unchanged after adaptations. In other words, we ensure the condition $A^*(p_1) = A(p_1)$ given the observed stagewise p-value p_1 regardless of an adaptation, where $A(p_1)$ and $A^*(p_1)$ are the conditional error functions before and after the adaptation. The dif-

ference between the Müller-Schäfer and the Denne methods is that Denne uses a particular $A(p_1)$ to obtain the closed form solution, while Müller and Shäfer consider a more general situation and emphasize that the $A(p_1)$ needs to be calculated only for the observed p_1 through simulations, and determination of the stopping boundaries afterwards is based on the condition $A^*(p_1) = A(p_1)$. They also stressed that this concept can be recursively used in a trial. We will discuss recursive approaches in the next chapter.

In this chapter, we have also discussed the conditional power for different designs. The conditional power can be used to compare different methods, and for the purpose of trial monitoring. When we force the treatment effect to zero, the conditional power function becomes the conditional error function, which allows for comparisons between different methods from the conditional error point of view.

Problem

7.1 Construct a table that contains stopping boundaries for each of the methods in Table 7.1 using (7.14).

7.2 Derive the conditional power for the two-stage adaptive design (see Exercise 4.2) with the test statistic:

$$T_k = \frac{1}{k}\sum_{i=1}^{k} p_i \text{ for k = 1 and 2,}$$

where p_i is the stagewise p-value based on subsample from the i^{th} stage.

Chapter 8

Recursive Adaptive Design

In this chapter, we will study the so-called recursive two-stage adaptive design (RTAD) (Chang, 2006). The recursive approach provides closed forms for stopping boundaries and adjusted p-values for any K-stage design and avoids any numerical integration; at the same time it allows for a broad range of adaptations such as SSR, dropping losers, and changing the number and timing of analyses without specification of an error-spending function. The key ideas of the RTAD are: (1) a K-stage design ($K > 1$) can be constructed using recursive two-stage designs; (2) the conditional error principle ensures that the recursive process will not inflate type-I error; and (3) the closed form solutions are obtained through recursively utilizing the two-stage design solutions for stopping boundary, adjusted p-value, and conditional power. In this approach, the trial is designed one step ahead at every interim analysis.

We will first introduce the concept of p-clud in Section 8.1 and review the two-stage approaches: MSP, MPP, and MINP in Section 8.2. In Section 8.3, we introduce the error-spending principle, from which we derive the conditional error principle. The later is a key element for deriving the recursive formula for the RTAD in Section 8.4. A clinical trial application of RTAD is also presented in Section 8.4. In Section 8.5 and 8.6, we will introduce two other adaptive methods for K-stage adaptive designs proposed by Müller and Shäfer (2004) and Brannath-Posch-Bauer (2002), respectively. Section 8.7 is a summary and discussion.

8.1 P-clud Distribution

The methods discussed in this chapter will assume the condition of p-clud (Brannath, Posch, and Bauer, 2002).

Definition 8.1 p-clud: P-value p_1 and p_2 are p-clud if the distribution

of p-value p_1 and the conditional distribution of p_2 given p_1 are stochastically larger than or equal to the uniform distribution on [0,1], i.e.,

$$\Pr_{H_o}(p_1 \leq \alpha) \leq \alpha \text{ and } \Pr_{H_o}(p_2 \leq \alpha | p_1) \leq \alpha, \forall \alpha \in [0,1]. \tag{8.1}$$

It is usually assumed that the p-values p_1 and p_2 are stochastically independent under H_o. Although stochastically independent sample units are recruited at the two stages, this is not necessarily the case. For an example, assume that rank tests are used at the two stages. The discrete distribution of p_2 under the null hypothesis may depend on p_1 via a sample-size reassessment rule: The experimenter may choose n_2 depending on the value observed for p_1. However, p_2 is still p-clud (Brannath, et al., 2002). However, when p_1 and p_2 are independent and uniformly distributed on [0,1], the level α is exhausted in (8.1).

We now study the distribution of stagewise p-values under the null hypothesis.

Let X be a continuous random variable with probability density function $f_x(x)$ and $F_X(x)$ be the c.d.f and $Y = g(X)$. $X = \eta(Y) = g^{-1}(Y)$ are monotonic increasing functions, where x is a realization of X, y is a realization of Y.

The p.d.f. of Y is given by (Kokoska and Zwillinger, 2000, p.40)

$$f_Y(y) = f_X(\eta(y))\eta'(y), \; g'(x) \neq 0. \tag{8.2}$$

From (8.2), we have

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{Y} f_Y(x)\,dx = \int_{-\infty}^{Y} f_X(\eta(x))\eta'(x)\,dx \\ &= \int_{-\infty}^{\eta(y)} f_X(\eta(x))\,d\eta(x). \end{aligned}$$

Therefore,

$$F_Y(y) = F_X(\eta(y)). \tag{8.3}$$

Now let X be the test statistic for testing $H_o : \delta = 0$ against the alternative $H_a : \delta > 0$. Let $(-\infty, y)$ be the rejection region, and $g(y) = F_Y(y|\delta)$, i.e., the p-value (p-value is reviewed as a function of critical point y), then $\eta(y) = F_Y^{-1}(y|\delta)$. Substituting this into (8.3), we obtain:

$$F_Y(y) = F_X\left(F_Y^{-1}(y|\delta)\right). \tag{8.4}$$

Notice that under the null hypothesis, $F_Y^{-1}(y|\delta = 0) = F_X^{-1}(y)$, (8.4) becomes

$$F_Y(y) = F_X\left(F_X^{-1}(y)\right) = y, \tag{8.5}$$

which implies that Y is uniformly distributed on [0,1].

8.2 Two-Stage Design

Consider a two-stage clinical trial in which a hypothesis test is performed at the interim analysis, followed by adaptations based on the interim results. Such adaptations can be sample-size adjustment, a change of the treatment allocation probabilities. The testing for efficacy of the experimental drug can be formulated using a global null hypothesis:

$$H_o : H_{o1} \cap H_{o2}, \tag{8.6}$$

where H_{oi} $(i = 1, 2)$ is the null hypothesis test at the i^{th} interim analysis. The stagewise p-value p_i is assumed to be uniformly distributed over [0,1] or p-clud. The test statistic for the k^{th} stage can be formulated using combinations of stagewise p_i, such as the product (Bauer and Kohne, 1994) $T_k = \Pi_{i=1}^k p_i,\ (k = 1, 2)$, the sum $T_k = \Sigma_{i=1}^k p_i,\ k = 1, 2$ (Chang 2006) and the inverse-normal transformation (Lehmacher and Wassmer,1999) $T_k = 1 - \Phi\left(\Sigma_{i=1}^k w_{ki}\Phi^{-1}(1 - p_i)\right),\ (k = 1, 2)$, where $\Sigma_{i=1}^k w_{ki}^2 = 1$. The stopping rules can be written as

$$\begin{cases} \text{Stop for efficacy if } T_k \leq \alpha_k\,, \\ \text{Stop for futility if } T_k > \beta_k, \\ \text{Continue, otherwise.} \end{cases} \tag{8.7}$$

For convenience, α_k and β_k are called the efficacy and futility boundaries, respectively.

A level-α test requires that

$$\alpha = \alpha_1 + \int_{\alpha_1}^{\beta_1} \int_{-\infty}^{\alpha_2} f_{T_1 T_2}\, dt_2 dt_1. \tag{8.8}$$

Because the p-value associated with a test is the smallest significance level α for which the null hypothesis is rejected (Robert, 1997, p.196), the stagewise-ordering p-value corresponding t, the realization of the test statistic (Jessinon and Turnbull, 2000, p.180 and 356, Chang 2006) is given by

$$p(t) = \alpha_1 + \int_{\alpha_1}^{\beta_1} \int_{-\infty}^{t} f_{T_1 T_2}\, dt_2 dt_1. \tag{8.9}$$

Note that the adjusted p-value is a measure of over statistical strength for rejecting H_o. The later the H_o is rejected, the larger the adjusted p-value is and the weaker the statistical evidence (against H_o) is. A late rejection leading to a larger p-value is reasonable because a portion of the alpha has been spent at the earlier stage. When the test statistic at the k^{th} stage $T_k = t = \alpha_k$ (i.e., just on the efficacy stopping boundary), the p-value is equal to the alpha spent up to the k^{th} stage.

8.2.1 *Method Based on Product of P-values*

For the method based on the product of the stagewise p-value (MPP), the test statistic is defined as

$$T_k = \Pi_{i=1}^{k} p_i,\ k = 1, 2. \tag{8.10}$$

The overall type-I error control requires that (Bauer and Kohne, 1994, Chang, 2006)

$$\alpha = \alpha_1 + \alpha_2 \ln \frac{\beta_1}{\alpha_1}. \tag{8.11}$$

The stagewise-ordering p-value corresponding to a test statistic t can be obtained by replacing α_k with t in (8.11) when the trial stops at the k^{th} stage, that is

$$p(t; k) = \begin{cases} t & \text{for } k = 1, \\ \alpha_1 + t \ln \frac{\beta_1}{\alpha_1} & \text{for } k = 2\ . \end{cases} \tag{8.12}$$

The conditional power is given by (Chapter 7)

$$P_c\left(p_1, \delta\right) = 1 - \Phi\left(\Phi^{-1}\left(1 - \frac{\alpha_2}{p_1}\right) - \frac{\delta}{\sigma}\sqrt{\frac{n_2}{2}}\right), \alpha_1 < p_1 \leq \beta_1. \tag{8.13}$$

The sample-size required for the second stage can be obtained by solving (8.13) for n_2 :

$$n_2 = \left[\frac{\sqrt{2}\sigma}{\delta}\left(\Phi^{-1}\left(1-\frac{\alpha_2}{p_1}\right)-\Phi^{-1}\left(1-P_c\right)\right)\right]^2. \tag{8.14}$$

The unconditional power is given by

$$Pw\left(\delta\right) = \int_{-\infty}^{+\infty} P_c\left(p_1,\delta\right) f_\delta\left(p_1\right) d\delta_1, \tag{8.15}$$

where $f_\delta\left(p_1\right)$ is p.d.f. of stagewise p-value p_1.

8.2.2 *Method Based on Sum of P-values*

For the method based on the sum of the stagewise p-values (MSP), the test statistic is defined as

$$T_k = \Sigma_{i=1}^{k} p_i,\ k = 1, 2. \tag{8.16}$$

A level α test requires that

$$\alpha = \alpha_1 + \alpha_2(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2),\ \alpha_2 \geq \beta_1. \tag{8.17}$$

By predetermining two of the three stopping boundaries α_1, β_1, and α_2, the third one can easily be obtained from (8.17). Note that for efficiency, we always choose $\beta_1 \leq \alpha_2$. This is because if $\beta_1 > p_1 > \alpha_2$, then $p_1 + p_2 > \alpha_2$ and H_o will definitely not be rejected. Hence there is no reason to continue the trial.

The adjusted p-value corresponding to a test statistic t can be obtained by replacing α_2 with t in (8.17), that is

$$p(t;k) = \begin{cases} t & \text{for } k = 1, \\ \alpha_1 + t(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2) & \text{for } k = 2\ . \end{cases} \tag{8.18}$$

The conditional power for MSP is given by

$$P_c\left(p_1,\delta\right) = 1 - \Phi\left(\Phi^{-1}\left(1-\alpha_2+p_1\right) - \tfrac{\delta}{\sigma}\sqrt{\tfrac{n_2}{2}}\right),\ \alpha_1 < p_1 \leq \beta_1. \tag{8.19}$$

The sample-size can be obtained by solving (8.19) for n_2 :

$$n_2 = \left[\frac{\sqrt{2}\sigma}{\delta} \left(\Phi^{-1}\left(1 - \alpha_2 + p_1\right) - \Phi^{-1}\left(1 - P_c\right) \right) \right]^2. \tag{8.20}$$

The unconditional power is given by

$$Pw\left(\delta\right) = \int_{-\infty}^{+\infty} P_c\left(p_1, \delta\right) f_\delta\left(p_1\right) d\delta_1, \tag{8.21}$$

where $f_\delta\left(p_1\right)$ is p.d.f. of stagewise p-value p_1.

8.2.3 *Method Based on Inverse-Normal P-values*

For the inverse-normal method, the test statistic is given by

$$\begin{cases} T_1 = p_1 \\ T_2 = 1 - \Phi\left(w_1 z_{1-p_1} + w_2 z_{1-p_2}\right) \end{cases}, \tag{8.22}$$

where $w_1^2 + w_2^2 = 1$.

Determination of the stopping boundaries and adjusted p-value requires numerical integrations or Monte Carlo simulations, which can be done by using SAS Macro 5.1 provided in Chapter 5. For an equal information design, the stopping boundaries are tabulated in Table 5.1.

The stagewise-ordering p-value corresponding to the observed $\mathrm{T}_k = t$ is given by

$$p(t;k) = \begin{cases} t & \text{for } k = 1, \\ \alpha_1 + pow\left(\alpha_2 = t\right) & \text{for } k = 2 \end{cases}, \tag{8.23}$$

where $Pow(\alpha_2 = t)$ is the probability of rejecting H_o at the second stage. The adjusted p-value can be obtained from SAS Macro 5.1 when set $\alpha_2 = t$.

The conditional power is given by

$$P_c\left(p_1, \delta\right) = 1 - \Phi\left(\frac{z_{1-\alpha_2} - w_1 z_{1-p_1}}{w_2} - \frac{\delta}{\sigma}\sqrt{\frac{n_2}{2}}\right), \alpha_1 < p_1 \leq \beta_1. \tag{8.24}$$

The sample-size is given by

$$n_2 = \left[\frac{\sqrt{2}\sigma}{\delta} \left(\frac{z_{1-\alpha_2} - w_1 z_{1-p_1}}{w_2} - z_{1-P_c} \right) \right]^2. \tag{8.25}$$

To visualize the differences of various methods (MIP, MSP, MPP, and MINP), the stopping boundaries at the second stage are plotted for the

same stopping boundaries ($\alpha_1 = 0.005$, $\beta_1 = 0.25$) at the first stage (Figure 8.1). We can see that MPP and MINP are similar. MSP and MIP has some constraints on the consistency of the results from the different stages. In other words, when p_1 and p_2 are very different, the statistical significance can not be declared using MSP or MIP. On the other hand, if $p_1 = 0.006$ and $p_2 = 0.6$ (100 times difference! The fact of one-sided $p_2 = 0.6$ indicates the wrong direction of the treatment effect!), the null H_o will still be rejected using MPP.

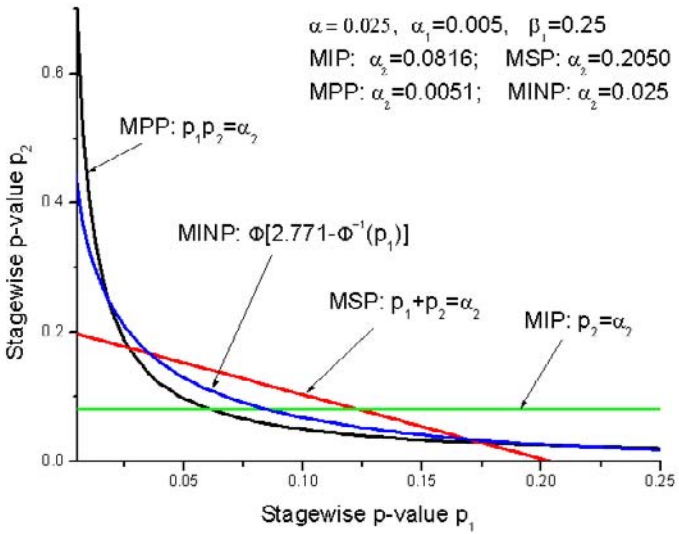


Figure 8.1: Various Stopping Boundaries at Stage 2

8.2.4 *Confidence Interval and Unbiased Median*

We now consider a general form of stagewise-ordering confidence intervals. Consider the null hypothesis

$$H_o : \delta = \delta_0. \tag{8.26}$$

In an adaptive design setting, (8.26) implies the global null hypothesis:$H_o = \cap_{i=1}^{K} H_{oi}$, where $H_{oi} : \delta = \delta_0$.

In general, a $100(1-\alpha)\%$ confidence interval consists of all δ_0 such that the null hypothesis (8.26) would not be rejected, given the observed value for the test statistic $T = t$.

It is obvious that the stagewise p-value for the k^{th} stage can be expressed as

$$\tilde{p}_k = 1 - \Phi\left(\frac{\Delta_k - \delta_0}{\hat{\sigma}}\sqrt{\frac{n_k}{2}}\right) , \tag{8.27}$$

where Δ_k = the observed mean difference $\Delta_k = \hat{\mu}_B - \hat{\mu}_A$ between the two groups for normal endpoint at the k^{th} stage, Δ_k = response rate difference $R_B - R_A$ between the two groups for binary variable at the k^{th} stage, and $\Delta_k = \hat{\lambda}_B - \hat{\lambda}_A$ hazard rate difference between the two groups for survival endpoint at the k^{th} stage. We have assumed a constant variance for large sample-size trial, which is given by (see Chapter 2)

$$\hat{\sigma}^2 = \begin{cases} \hat{\sigma}^2 & \text{for normal endpoint,} \\ R_o(1-R_o) & \text{for binary endpoint,} \\ \lambda_o^2\left[1 - \frac{e^{\lambda_o T_0}-1}{T_0\lambda_o e^{\lambda_o T_s}}\right]^{-1} & \text{for survival endpoint,} \end{cases} \tag{8.28}$$

where R_0 = response rate under H_o and λ_o is hazard rate under H_o. They can be replaced with the observed pooled value.

Because $\frac{\Delta_k}{\hat{\sigma}}\sqrt{\frac{n_k}{2}} = z_{1-p_k}$, where p_k is the stagewise p-value for the k^{th} stage for the null hypothsis (8.26) with $\delta_0 = 0$, (8.27) can be written as

$$\tilde{p}_k = 1 - \Phi\left(z_{1-p_k} - \frac{\delta_0}{\hat{\sigma}}\sqrt{\frac{n_k}{2}}\right) , \tag{8.29}$$

or

$$\tilde{p}_k = \Phi\left(\frac{\delta_0}{\hat{\sigma}}\sqrt{\frac{n_k}{2}} - z_{1-p_k}\right) . \tag{8.30}$$

We now construct the test statistic for (8.26) using MPP, MSP, and MINP as follows.

For MPP, MSP, and MINP, if the trial is stopped at the first stage, the confidence interval bound δ_{01} is given by

$$\delta_{01} = \hat{\sigma}\sqrt{\frac{2}{n_1}}\left(\Phi^{-1}(1-\alpha_1) + z_{1-p_1}\right). \tag{8.31}$$

When the trial is stopped at the second stage, the confidence bound calculations are different for MPP, MSP, and MINP as discussed below.

(1) For MPP, the test statistic is given by

$$T_2 = \prod_{i=1}^{2} \Phi\left(\frac{\delta}{\hat{\sigma}}\sqrt{\frac{n_i}{2}} - z_{1-p_i}\right).$$

To obtain the confidence limit δ_{02} for δ, let $T_2 = \alpha_2$ and we obtain:

$$\prod_{i=1}^{2} \Phi\left(\frac{\delta_{02}}{\hat{\sigma}}\sqrt{\frac{n_i}{2}} - z_{1-p_i}\right) = \alpha_2. \tag{8.32}$$

(8.32) can be solved numerically for δ_{02}.

(2) For MSP, the test statistic is given by

$$T_2 = \sum_{i=1}^{2} \Phi\left(\frac{\delta}{\hat{\sigma}}\sqrt{\frac{n_i}{2}} - z_{1-p_i}\right).$$

To obtain the confidence limit δ_{02} for δ, let $T_2 = \alpha_2$ and we obtain:

$$\sum_{i=1}^{2} \Phi\left(\frac{\delta_{02}}{\hat{\sigma}}\sqrt{\frac{n_i}{2}} - z_{1-p_i}\right) = \alpha_2. \tag{8.33}$$

We can numerically solve (8.33) for δ_{02}.

(3) For MINP, the test statistic test is given

$$T_2 = \sum_{i=1}^{2} \Phi\left(\frac{\delta}{\hat{\sigma}}\sqrt{\frac{n_i}{2}} - z_{1-p_i}\right).$$

To obtain the confidence limit δ_{02} for δ, let $T_2 = \alpha_2$ and we obtain:

$$1 - \Phi\left(\sum_{i=1}^{2} w_i\left(z_{1-p_i} - \frac{\delta_{02}}{\hat{\sigma}}\sqrt{\frac{n_i}{2}}\right)\right) = \alpha_2. \tag{8.34}$$

(8.34) can be solved analytically for δ_{02}, i.e.,

$$\delta_{02} = \hat{\sigma}\sqrt{2}\frac{w_1 z_{1-p_1} + w_2 z_{1-p_2} - \Phi^{-1}(1-\alpha_2)}{w_1\sqrt{n_1} + w_2\sqrt{n_2}}, \tag{8.35}$$

which is consistent with the result given by (3.31) in Chapter 3 with $k = 2$.

We now can summarize the steps for calculating the confidence intervals as follows:

(1) Using (8.30) and (8.32), (8.33) or (8.35) to calculate δ_{0k} dependent on MPP, MSP, or MINP, respectively.

(2) The one-sided $(1-\alpha)\%$ overall confidence bound is given by $\delta_c = \max\{\delta_{01}, \delta_{02}\}$.

(3) The one-sided $(1-\Sigma_{i=1}^{k}\pi_i)\%$ stagewise-ordering confidence interval bound is given by

$$\delta_c = \max\{\delta_{01}, ...\delta_{0k}\},$$

where k is the stage at which the null was actually rejected.

(4) The 50% confidence interval bound gives unbiased median estimate.

Note that we have assumed that we will not change the adaptive design with regard to the number and timing of the analyses and the stopping boundaries α_k regardless of whether we believe $\delta = \delta_0$ or $\delta = 0$.

If the method to calculate the stagewise p-values is different from the above methods, the calculations of confidence interval and point estimate are similar and can be described as the following steps:

(1) Calculate the stagewise p-value for the k^{th} stage as a function of a selected value for δ_{0k} and the observed data $\hat{\mathbf{X}}_k$, i.e., $p_k\left(\hat{\mathbf{X}}_k, \delta_{ok}\right)$.

(2) Calculate the test statistic T_k based on (8.10), (8.16), or (8.22) dependent on MPP, MSP, or MINP, respectively.

(3) If $T_k > \alpha_k$, then reduce δ_{0k}; if $T_k < \alpha_k$, then increase δ_{0k}. Go back to step 1, using the new value for δ_{0k}.

(4)Continue the process until $T_k = \alpha_k$, or they are close enough.

(5) The final δ_{0k}, which is the one-sided $(1-\Sigma_{i=1}^{k}\pi_i)\%$ stagewise confidence interval bound given the null is rejected at the stage k.

(6) The one-sided $(1-\alpha)\%$ overall confidence bound is given by

$$\delta_c = \max\{\delta_{01}, \delta_{02}\}.$$

(7) The one-sided $(1-\Sigma_{i=1}^{k}\pi_i)\%$ stagewise-ordering confidence interval bound is given by

$$\delta_s = \max\{\delta_{i1}, ...\delta_{ik}\},$$

where k is the stage at which the null was actually rejected.

(8) The 50% confidence interval bound δ_k gives unbiased median estimate.

8.3 Error-Spending and Conditional Error Principles

We are going to formally introduce the so-called error-spending principle, from which we derive the second useful principle: the conditional error principle. The latter will be used to construct the recursive two-stage adaptive designs in the next section. The error-spending principle has been implicitly used from time to time (Lan and DeMets, 1983). The conditional error principle also appears informally in a different form (Müller and Shäfer, 2001).

In an initial K-stage adaptive trial for testing the global null hypothesis $H_o : \cap_{j=1}^{K} H_{oj}$, where H_{oj} is the null hypothesis at j^{th} stage, the type-I error control requirement can be expressed as

$$\alpha = \pi_1 + ... + \pi_k + \pi_{k+1} + ... + \pi_K, \tag{8.36}$$

where π_i is α spent at the i^{th} stage in the initial design of an adaptive trial.

Now assume that after the k^{th} interim analysis, adaptations are made that result in changes to stagewise hypotheses and error spending after the k^{th} stage. We denote H_{oi}^* and π_i^* $(i = k+1, ..., K^*)$ as the new hypotheses and the new error spending, respectively, where K^* is the new total number of analyses. The overall α-control becomes

$$\alpha = \pi_1 + ... + \pi_k + \pi_{k+1}^* + ... + \pi_{K^*}^*. \tag{8.37}$$

From (8.37), the unconditional error rate after the k^{th} stage can be expressed as

$$\alpha - \sum_{j=1}^{k} \pi_j = \sum_{j=k+1}^{K} \pi_j = \int_{\alpha_k}^{\beta_k} A\left(p_k\right) dp_k. \tag{8.38}$$

Similarly, from (8.38), the unconditional error rate after the k^{th} stage can be expressed as

$$\alpha - \sum_{j=1}^{k} \pi_j = \sum_{j=k+1}^{K^*} \pi_j^* = \int_{\alpha_k}^{\beta_k} A^*\left(p_k\right) dp_k, \tag{8.39}$$

where $A\left(p_k\right)$ and $A^*\left(p_k\right)$ are called conditional error functions (under the global null hypothesis). α_k and β_k are constants. Comparing (8.38) and (8.39), we immediately have the following error-spending principle.

Error-spending principle: In an adaptive trial, if an adaptation at k^{th} stage ensures the invariance of unconditional error, i.e.,

$$\int_{\alpha_k}^{\beta_k} A\left(p_k\right) dp_k = \int_{\alpha_k}^{\beta_k} A^*\left(p_k\right) dp_k, \tag{8.40}$$

then the overall α is controlled under the global null hypothesis:

$$H_o : \left(\cap_{j=1}^{k} H_{oj}\right) \cap \left(\cap_{j=k+1}^{K^*} H_{oj}^*\right). \tag{8.41}$$

The principle (8.40) is very general. The commonly used error-spending approach (Lan-DeMets, 1983) is a special form of the error-spending principle, which requires prespecification of the unconditional error as a function of information time. The conditional error method (Proschan and Hunsberger, 1995) is a special use of the error-spending principle for two-stage designs. Importantly, if we let $A\left(p_k\right) = A^*\left(p_k\right)$, then (8.40) holds and the overall α is controlled. Therefore we can formally introduce the following principle.

Conditional error principle: In an adaptive trial, if an adaptation at the k^{th} stage ensures the invariance of conditional error, i.e.,

$$A^*\left(p_k\right) = A\left(p_k\right), \tag{8.42}$$

then the overall α is controlled under the global null hypothesis (8.41).

The conditional error principle allows changes in the total number of analyses, the timing of analyses, randomization, hypothesis changes, etc. in adaptive trials. The principle proposed by Müller and Shäfer (2001) is the applied Brownian model at pre-defined time points of the interim analyses (Müller and Shäfer, 2004). Müller and Shäfer (2004) extend their principle and allow for changes in the timing of an interim analysis. They did not explicitly specify the condition (8.41). The significance of (8.42) is that it provides a simple but very general way to control α by simulating the conditional error on flying; at same time, it allows for different adaptations. (8.42) tells us that one can make any adaptation as long as the conditional error is unchanged. (8.42) can be used repeatedly, meaning that one can apply adaptations as many times as one wants as long as the conditional error rate remains unchanged at each stage.

The conditional error function for MPP, MSP, and MINP can be obtained by substituting $\delta = 0$ into the conditional power formula, i.e., (8.13), (8.19), and (8.24), respectively, and incorporating the corresponding $B(\alpha_2, p_1)$:

$$A(p_1) = \frac{\alpha_2}{p_1},\ \alpha_1 < p_1 \leq \beta_1 \tag{8.43}$$

$$A(p_1) = \alpha_2 - p_1,\ \alpha_1 < p_1 \leq \beta_1, \tag{8.44}$$

and

$$P_c(p_1, \delta) = 1 - \Phi\left(\frac{z_{1-\alpha_2} - w_1 z_{1-p_1}}{w_2}\right),\ \alpha_1 < p_1 \leq \beta_1. \tag{8.45}$$

8.4 Recursive Two-Stage Design

We can see that the conditional error principle allows for a broad range of adaptations without inflating the alpha, but the calculation of conditional error usually requires computer simulations as indicated by Müller and Shäfer (2001). In this section, we will derive a closed form solution for a K-stage design that allows for a broad range of adaptations by recursively using the conditional error principle and two-stage formula for the conditional error, stopping boundary, adjusted p-value, and confidence interval. Naturally, this method is called recursive two-stage adaptive design (RTAD).

First, let's explain the concept and mechanics of the recursive two-stage design. Suppose we want to design an adaptive trial, but we may not know how many stages and what allowable adaptations will be best to meet the trial objectives. We start with a two-stage design. At the interim look, we can make adaptations including SSR, increasing the number of analyses, etc., but we don't want to fully specify the rules. To meet the requirement, we can make the "short-term" plan by adding one more analysis into the design. Now the trial became a 3-stage adaptive design. However, instead of constructing a statistical method based on a 3-stage design, we view this 3-stage design as a stagnation of 2, two-stage designs. In other words, stages 1 and 2 are considered the first two-stage design, and stages 2 and 3 are viewed as the second two-stage design. In general, the k^{th} and $(k+1)^{th}$ stages are considered to be the k^{th} two-stage design. Each new two-stage design is tested at a different level of α that is equal to the newly calculated conditional error rate (Figure 8.2).

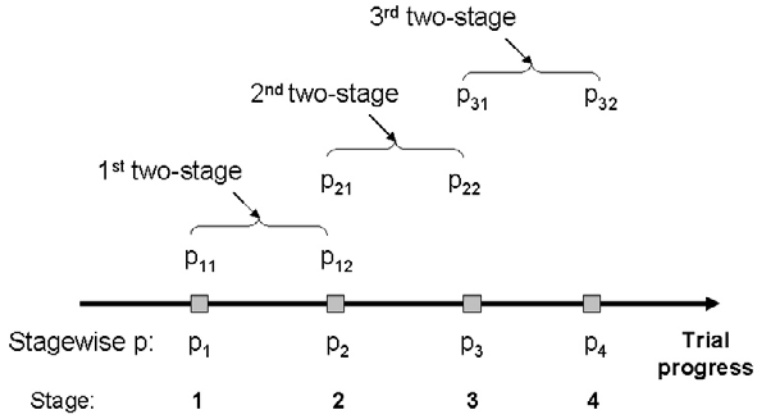


Figure 8.2: Recursive Two-stage Adaptive Design

In what follows, we will derive the formulation for recursive two-stage adaptive design based on MSP and MPP. The reason we can derive the closed form is that we have the explicit forms of the conditional error probability and stopping boundaries for two-stage designs.

8.4.1 *Sum of Stagewise P-values*

Suppose at a typical stage i, we decide to increase the number of stages in addition to other adaptations. The test statistic for the i^{th} two-stage design is defined as

$$\begin{cases} T_{i1} = p_{i1} & \text{for stage 1,} \\ T_{i2} = p_{i1} + p_{i2} & \text{for stage 2,} \end{cases} \tag{8.46}$$

where $p_{i1} = p_i$ and $p_{i2} = p_{i+1}$, p_i is the naive stagewise p-value based on the subsample from the i^{th} stage. Note that p_{i1} and p_{i2} are mutually independent and uniformly distributed over [0, 1] under H_o.

If $p_{i1} > \beta_{i1}$, stop the trial and accept H_o; if $p_{i1} \leq \alpha_{i1}$, stop the trial and reject H_o;if $\alpha_{i1} < p_{i1} \leq \beta_{i1}$, the trial continues and we can either go with the previous plan with the stopping boundary α_{i2} for the final stage with or without sample-size adjustment, or plan the next "two-stage" design. Based on the conditional error principle, the new two-stage design should be tested at level $A(p_{i1})$, where $A(p_{i1})$ is the conditional error rate at the i^{th} stage. From (8.28), we can obtain the conditional error rate at the i^{th} IA (we always choose $\beta_{i1} < \alpha_{i2}$):

$$A\left(p_{i1}\right) = \alpha_{i2} - p_{i1},\ \alpha_{i1} < p_{i1} \leq \beta_{i1}. \tag{8.47}$$

The stopping boundaries are determined for the new two-stage design by

$$A\left(p_{i1}\right) = \alpha_{i+1,1} + \alpha_{i+1,2}(\beta_{i+1,1} - \alpha_{i+1,1}) - \frac{1}{2}(\beta_{i+1,1}^2 - \alpha_{i+1,1}^2),\ i = 0, 1, ... \tag{8.48}$$

where for convenience, we define $A\left(p_{01}\right) = \alpha$. We can predetermine $\beta_{i+1,1}$ and $\alpha_{i+1,1}$, then the third value is given by

$$\alpha_{i+1,2} = \frac{A\left(p_{i1}\right) + \frac{1}{2}(\beta_{i+1,1}^2 - \alpha_{i+1,1}^2) - \alpha_{i+1,1}}{\beta_{i+1,1} - \alpha_{i+1,1}}. \tag{8.49}$$

Conditional power is similar to (8.19), but replace α_2 with α_{i2} and p_1 with p_{i1}; that is,

$$P_c\left(p_{i1}, \delta\right) = 1 - \Phi\left(\Phi^{-1}\left(1 - \alpha_{i2} + p_{i1}\right) - \frac{\delta}{\sigma}\sqrt{\frac{n_{i2}}{2}}\right),\ \alpha_{i1} < p_{i1} \leq \beta_{i1}. \tag{8.50}$$

Similar to (8.20), the sample-size based on the target conditional power P_c is given by

$$n_{i2} = \left[\frac{\sqrt{2}\sigma}{\delta}\left(\Phi^{-1}\left(1 - \alpha_{i2} + p_{i1}\right) - \Phi^{-1}\left(1 - P_c\right)\right)\right]^2. \tag{8.51}$$

We now consider stagewise-ordering adjusted p-values using the recursive approach following Brannath, Posch, and Bauer's idea. We review the collection of the stages after the 1^{st} stage of the i^{th} two-stage design as the second stage of the i^{th} design for the purpose p-value calculations. Of course this is just approximation. Similar to (8.18), we can define the p-value for the second stage of the $(i-1)^{th}$ two-stage design using backwards recursion as

$$\begin{cases} p_{K_0-1,2} = \begin{cases} t & \text{for } k = 1, \\ \alpha_{K_0 1} + t(\beta_{K_0 1} - \alpha_{K_0 1}) - \frac{1}{2}(\beta_{K_0 1}^2 - \alpha_{K_0 1}^2) & \text{for } k = 2\ . \end{cases} \\ p_{i-1,2} = \alpha_{i1} + \left(p_{i1} + p_{i,2}\right)(\beta_{i1} - \alpha_{i1}) - \frac{1}{2}(\beta_{i1}^2 - \alpha_{i1}^2),\ \text{for } i = 1, ..., K_0 - 1\ . \end{cases} \tag{8.52}$$

Then $p_{0,2}$ is the stagewise-ordering p-value.

8.4.2 *Product of Stagewise P-values*

For a test statistic based on the product of stagewise p-values,

$$\begin{cases} T_{i1} = p_{i1} & \text{for stage 1,} \\ T_{i2} = p_{i1}p_{i2} & \text{for stage 2.} \end{cases} \tag{8.53}$$

Similar to (8.13), the conditional error rate at the i^{th} IA is given by

$$A\left(p_{i1}\right) = 1 - \frac{\alpha_{i2}}{p_{i1}},\ \alpha_{i1} < p_{i1} \leq \beta_{i1}. \tag{8.54}$$

If $p_{i1} > \beta_{i1}$, stop the trial and accept H_o; if $p_{i1} \leq \alpha_{i1}$, stop the trial and reject H_o; if $\alpha_{i1} < p_{i1} \leq \beta_{i1}$, the trial continues and we can either go with the previous plan with the stopping boundary α_{i2} for the final stage or plan the next "two-stage" design at level $A\left(p_{i1}\right)$. For the next two-stage design, we can predetermine $\beta_{i+1,1}$ and $\alpha_{i+1,1}$, then the third value is given by

$$\alpha_{i+1,2} = \frac{A\left(p_{i1}\right) - \alpha_{i+1,1}}{\ln \beta_{i+1,1} - \ln \alpha_{i+1,1}}. \tag{8.55}$$

If the trial stops at the k^{th} stage, the stagewise-ordering adjusted p-value is approximately calculated using the backwards recursion:

$$\begin{cases} p_{K_0-1,2} = \begin{cases} t & \text{for } k = 1, \\ \alpha_{K_0 1} + t \ln \frac{\beta_{K_0 1}}{\alpha_{K_0 1}} & \text{for } k = 2\ . \end{cases} \\ p_{i-1,2} = \alpha_{i1} + \left(p_{i1} + p_{i,2}\right) \ln \frac{\beta_{i1}}{\alpha_{i1}},\ \text{for } i = 1, ..., K_0 - 1\ . \end{cases} \tag{8.56}$$

Then $p_{0,2}$ is the stagewise-ordering p-value.

8.4.3 *Inverse-Normal Stagewise P-values*

For the method based on inverse-normal stagewise p-value (MINP),

$$\begin{cases} T_{i1} = 1 - z_{1-p_{i1}} & \text{for stage 1,} \\ T_{i2} = 1 - \Phi\left(w_1 z_{1-p_{i1}} + w_2 z_{1-p_{i2}}\right) & \text{for stage 2.} \end{cases} \tag{8.57}$$

Similar to (8.27), the conditional error rate at the i^{th} IA is given by

$$A\left(p_{i1}\right) = 1 - \Phi\left(\frac{z_{1-\alpha_{i2}} - w_1 z_{1-p_{i1}}}{w_2}\right),\ \alpha_{i1} < p_{i1} \leq \beta_{i1}. \tag{8.58}$$

If $p_{i1} > \beta_{i1}$, stop the trial and accept H_o; if $p_{i1} \leq \alpha_{i1}$, stop the trial and reject H_o;if $\alpha_{i1} < p_{i1} \leq \beta_{i1}$, the trial continues and we can either go with the previous plan with the stopping boundary α_{i2} for the final stage or plan the next "two-stage" design at level $A(p_{i1})$. Again the stop boundaries can be determined through simulations using SAS Macro 5.1.

If the trial stops at the k^{th} stage, the stagewise-ordering adjusted p-value is calculated using the backwards recursion:

$$\begin{cases} p_{K_0-1,2} = \begin{cases} t & \text{for } k=1, \\ pow\left(\alpha_{K_0-1,2} = t\right) & \text{for } k=2 \end{cases}, \\ p_{i-1,2} = pow\left(\alpha_{i2} = (p_{i1} + p_{i,2})\right) \text{ for } k=2. \end{cases} \tag{8.59}$$

Again, SAS Macro 5.1 can be used to calculate $pow(\cdot)$.

8.4.4 *Confidence Interval and Unbiased Median*

For MPP, MSP, and MINP, if the trial is stopped at the first stage of the i^{th} two-stage design, the confidence interval bound δ_{i1} is given by

$$\delta_{i1} = \hat{\sigma}\sqrt{\frac{2}{n_{i1}}}\left(\Phi^{-1}(1-\alpha_{i1}) + z_{1-p_{i1}}\right) \text{ for } i = 1, ...K_0, \tag{8.60}$$

where K_0 is the total number of two-stage design.

When the trial is stopped at the second stage of the K_0 two-stage design, the confidence bound calculations for the last stage are different for MPP, MSP, and MINP as discussed below.

(1) For MPP, the confidence limit $\delta_{K_0 2}$ can be obtained by solving (8.61) numerically:

$$\Phi\left(\frac{\delta_{K_0 2}}{\hat{\sigma}}\sqrt{\frac{n_{K_0 1}}{2}} - z_{1-p_{K_0,1}}\right)\Phi\left(\frac{\delta_{K_0 2}}{\hat{\sigma}}\sqrt{\frac{n_{K_0 2}}{2}} - z_{1-p_{K_0,2}}\right) = \alpha_{K_0 2}. \tag{8.61}$$

(2) For MSP, the confidence limit δ_{i2} can be obtained by solving (8.62) numerically:

$$\Phi\left(\frac{\delta_{K_0 2}}{\hat{\sigma}}\sqrt{\frac{n_{K_0\,1}}{2}} - z_{1-p_{K_0 1}}\right) + \Phi\left(\frac{\delta_{K_0 2}}{\hat{\sigma}}\sqrt{\frac{n_{K_0 2}}{2}} - z_{1-p_{K_0,\ 2}}\right) = \alpha_{K_0 2}. \tag{8.62}$$

(3) For MINP, the confidence limit is given by

$$\delta_{i2} = \hat{\sigma}\sqrt{2}\frac{w_{K_0 1}z_{1-p_{K_0 1}} + w_{K_0 2}z_{1-p_{K_0,\ 2}} - \Phi^{-1}(1-\alpha_{K_0 2})}{w_{K_0 1}\sqrt{n_1} + w_{K_0 2}\sqrt{n_2}}. \tag{8.63}$$

The one-sided $(1-\alpha)\%$ overall confidence bound is given by

$$\delta_c = \max_{1 \le i \le K_0 - 1} \left\{ \delta_{i1}, \delta_{K_0\, k} \right\}, \tag{8.64}$$

where K_0 is the total number of two-stage designs.

The one-sided $(1 - \Sigma_{i=1}^{k} \pi_i)\%$ stagewise-ordering confidence interval bound is given by

$$\delta_s = \max_{1 \le i \le k} \left\{ \delta_{i1}, ...\delta_{ik} \right\}, \tag{8.65}$$

where k is the stage at which the null was actually rejected.

The 50% confidence interval bound δ_{ks} gives unbiased median estimate, where k is the stage where the trial was actually stopped.

8.4.5 *Application Example*

Example 8.1 Recursive Two-Stage Adaptive Design

For an adaptive design, the conditional power is often more important than the unconditional power because the sample-size can be adjusted to reach the desired (conditional) power. In other words, we can design a trial with low power, but allow for SSR later. A recursive two-stage design is ideal for this kind of adaptive design.

Suppose we are planning an early acute coronary syndrome (ACS) trial with a composite endpoint of death and myocardial infarction (MI) within 30 days of treatment. The event rate is 11% for the control group and 13% for the test group. A classic two-group design requires 5546/group to have 90% power at a significance level of 2.5% (one-sided test). However, the estimation of 11% is an approximation. To reduce the risk, interim analyses are considered. In case there is a very small treatment difference, the trial will allow for early stopping; if the effect size is moderate we increase the sample-size; if the effect size is big, we want to have a chance to make an earlier efficacy claim. Consider the uncertainty of the recruitment rate and inflexibility of the IDMC board's schedule. The information times for interim analyses may deviate from the plan. One way to deal with this is to use the error-spending approach to adjust the stagewise alpha according to the information time. However, the error spending approach requires predetermination of the error-spending function, which may lead to undesirable operating characteristics. Therefore, it is desirable to have the flexibility to adjust both the timing and error-spending independently. We are going to illustrate how to use the recursive two-stage design to accomplish this.

We think that 2 interim analyses and one final analysis will be a reasonable way of reducing risk and potentially shorten the time to market. Also, this plan is operationally feasible. Therefore, we start the first two-stage design with 4000 patients per group; the first stage will have 2000 patients per group. The first interim analysis is used to adjust the sample-size and assess futility stopping, but not efficacy stopping. The algorithm for SSR does not have to be specified at this moment. We use MSP with stopping boundaries: one-sided $\alpha_{11} = 0$ and $\beta_{11} = 0.2$; then $\alpha_{12} = 0.2250$ is calculated from (8.49). The power of rejecting the null hypothesis of no treatment effect at the interim analysis is zero and about 78% at the second stage. This power is calculated approximately from a classic design with 4000/group. The exact power can be calculated using simulation, but we don't have to do that because the conditional power is more important and can be easily calculated using (8.50) at the time of the interim analysis.

At the first IA, suppose we observe event rates $r_1 = 0.129$ and $r_2 = 0.114$ for the control and test groups, respectively. We estimate a treatment difference $\hat{\delta} = r_1 - r_2 = 0.015$, and standard deviation $\hat{\sigma} = \sqrt{[r_1(1-r_1) + r_2(1-r_2)]/2} = 0.3266$. The chi-square test statistic or equivalently, the z-score is $z_1 = \frac{\delta}{\hat{\sigma}}\sqrt{n_1/2} = 1.452$ and the corresponding stagewise-ordering one-sided p-value $p_{11} = 1 - \Phi(1.452) = 0.0732 < \beta_{11}$. Therefore the trial should continue. We assume that $\delta = \hat{\delta}$, $\sigma = \hat{\sigma}$. To reach a conditional power of $P_c = 90\%$, the sample-size is calculated from (8.50), i.e.,

$$\begin{aligned} n_{12} &= \left[\frac{\sqrt{2}\sigma}{\delta}\left(\Phi^{-1}(1-\alpha_{12}+p_{11}) - \Phi^{-1}(1-P_c)\right)\right]^2 \\ &= \left[\frac{0.3266\sqrt{2}}{0.015}\left(\Phi^{-1}(1-0.225+0.0732) - \Phi^{-1}(1-0.9)\right)\right]^2 \\ &= [30.792(1.0287+1.2814)]^2 \\ &= 5060. \end{aligned}$$

We don't want to simply increase the sample-size and delay the analysis because timing is so important in this trial. Therefore, we construct the second two-stage design as specified below:

(1) calculate the conditional error using (8.47): $A(p_{11}) = \alpha_{12} - p_{11} = 0.225 - 0.0732 = 0.151\,8$;

(2) choose $\alpha_{21} = 0.1$ and $\beta_{21} = 0.4$, (8.48) can be written as

$$0.1518 = 0.1 + \alpha_{22}(0.4 - 0.1) - \frac{1}{2}\left(0.4^2 - 0.1^2\right).$$

Solving it for α_{22}, we obtain $\alpha_{22} = 0.42267$.

(3) select $n_{21} = 2000$/group and $n_{22} = 3000$/group for the second two-stage design. Note that α_{21} and n_{21} are chosen such that the conditional power (assume $\delta = \hat{\delta} = 0.015$) for rejecting H_o at the 2nd IA (i.e., the first stage of the second two-stage design) is reasonably high, specifically,

$$\begin{aligned} P_c\left(p_{11}, \delta\right) &= 1 - \Phi\left(\Phi^{-1}\left(1 - \alpha_{12} + p_{11}\right) - \frac{\delta}{\sigma}\sqrt{\frac{n_{12}}{2}}\right) \\ &= 1 - \Phi\left(1.0287 - \frac{0.015}{0.3266}\sqrt{\frac{2000}{2}}\right) \\ &= 1 - \Phi\left(-0.423\,67\right) = 66.41\%. \end{aligned}$$

The sample-size n_{22} is not important at all because we can either change it later or add a new two-stage design. Suppose at the second IA, the event rate is $r_1 = 0.129$ and $r_2 = 0.116$. The stagewise p-value $p_{21} = p_2 = 1 - \Phi\left(\frac{0.013}{0.3278}\sqrt{\frac{2000}{2}}\right) = 0.1049 > \alpha_{21} = 0.1$ and H_o should not be rejected.

We may be curious about the classic design with the same data. In fact, if we pool 4000 subjects per group from stages 1 and 2 to calculate the p-value, as classic design the one-sided p-value will be equal to $\Phi\left(-\frac{(0.015+0.013)/2}{\sqrt{(0.3266^2+0.3278^2)/2}}\sqrt{\frac{4000}{2}}\right) = \Phi\left(-1.913\,5\right) > 0.0278 > \alpha = 0.025$. Therefore the classic design would fail to reject the null.

We now design the third two-stage design:

(1) calculate the conditional error using (8.47): $A\left(p_{21}\right) = \alpha_{22} - p_{21} = 0.42267 - 0.1049 = 0.317\,77$;

(2) choose $\alpha_{31} = 0.2$ and $\beta_{31} = 0.6$,

$$0.31777 = 0.2 + \alpha_{32}(0.6 - 0.2) - \frac{1}{2}\left(0.6^2 - 0.2^2\right).$$

Solve it, we have $\alpha_{32} = 0.694\,43$.

(3) select $n_{21} = 2000$/group and $n_{22} = 2000$/group for the second two-stage design. Assuming the true parameters are the estimates from the pooled data: $\delta = (0.015 + 0.013)/2 = 0.014$ and $\sigma = \sqrt{\left(0.3266^2 + 0.3278^2\right)/2} = 0.3272$. The conditional power is calculated as

$$\begin{aligned} P_c\left(p_{21}, \delta\right) &= 1-\Phi\left(\Phi^{-1}\left(1-\alpha_{22}+p_{21}\right)-\frac{\delta}{\sigma}\sqrt{\frac{n_{22}}{2}}\right) \\ &= 1-\Phi\left(0.783\,4-\frac{0.014}{0.3272}\sqrt{\frac{2000}{2}}\right) \\ &= 71.55\%. \end{aligned}$$

Suppose at the first stage IA of the third two-stage design (i.e., the third stage), the event rate gain is $r_1 = 0.129$ and $r_2 = 0.116$. The stagewise p-value $p_{31} = p_3 = 1-\Phi\left(\frac{0.013}{0.3278}\sqrt{\frac{2000}{2}}\right) = 0.1049 < \alpha_{21} = 0.2$ and H_o is rejected.

Note if we reestimate the sample-size for the second stage a little bit higher conditional power, say, 58.19% instead of 56.78%, the required sample-size is $n_{21} = 2100$, $p_{21} = 0.0994 < \alpha_{21}$ and the H_o would be rejected at the second stage.

We can calculate the adjusted p-value and point and confidence interval:

(1) The pooled standard deviation

$$\sigma = \sqrt{\left(0.3266^2 + 0.3278^2 + 0.3278^2\right)/3} = 0.327\,4.$$

(2) Because the trial is actually stopped at the first stage of the third two-stage ($K_0 = 3$) design, The adjusted stagewise-ordering p-value can be calculated using (8.52), i.e.,

$$\begin{cases} p_{2,2} = t = p_{31} = 0.1049 \\ p_{1,2} = \alpha_{21} + \left(p_{21} + p_{2,2}\right)\left(\beta_{21} - \alpha_{21}\right) - \frac{1}{2}(\beta_{21}^2 - \alpha_{21}^2) \\ = 0.2 + (0.0994 + 0.1049)(0.4 - 0.1) - \frac{1}{2}\left(0.4^2 - 0.1^2\right) \\ = 0.186\,29 \\ p_{0,2} = \alpha_{11} + \left(p_{11} + p_{1,2}\right)\left(\beta_{11} - \alpha_{11}\right) - \frac{1}{2}(\beta_{11}^2 - \alpha_{11}^2) \\ = 0 + (0.0732 + 0.186\,29)\left(0.2 - 0\right) - \frac{1}{2}\left(0.2^2 - 0^2\right) \\ = 0.0319. \end{cases}$$

We may be curious about the p-value from a three-stage MINP, which is given by

$$\begin{aligned} p &= 1-\Phi\left(\frac{1}{\sqrt{K}}\sum_{i=1}^{K}\Phi^{-1}\left(1-p_i\right)\right) \\ &= 1-\Phi^{-1}(1.254 + 1.285 + 1.4522)/\sqrt{3} \\ &= 0.0106. \end{aligned}$$

The summary of the recursive design are presented in Table 8.1.

Table 8.1: Summary of the Recursive Two-Stage Design

α_{11}	β_{11}	α_{12}	α_{21}	β_{21}	α_{22}	α_{31}	β_{31}	α_{32}
0	0.200	0.225	0.100	0.400	0.423	0.200	0.600	0.694
r_1	r_2	σ	r_1	r_2	σ	r_1	r_2	σ
0.129	0.114	0.327	0.129	0.116	0.328	0.129	0.116	0.328
$z_{1-p_{11}}$	p_{11}	α	$z_{1-p_{21}}$	p_{21}	$\mathrm{A}(p_{11})$	$z_{1-p_{31}}$	p_{31}	$\mathrm{A}(p_{21})$
1.452	0.073	0.025	1.029	0.105	0.152	1.029	0.105	0.312

Note that $p_{i,2}$ $(i < K_0)$ is not exactly calculated from the observed data in the sample as calculating p_{i1}. Therefore, $\mathrm{p}_{0,2}$ can not be used in comparison to the one-sided α for rejection.

(3) Confidence limits of the rate difference can be calculated using the approximations (8.60) (we don't need (8.62) because the trial is stopped at the first stage of the 3rd two-stage design):

$$\delta_{11} = 0.3266\sqrt{\frac{2}{2000}}\,(0.5 + 1.4522) = 0.02016,$$

$$\delta_{21} = 0.3278\sqrt{\frac{2}{2000}}\,(-1.2814 + 1.0286) = -0.00262,$$

$$\delta_{31} = 0.3278\sqrt{\frac{2}{2000}}\,(-0.8415 + 1.0286) = 0.001\,92.$$

8.5 Recursive Combination Tests

In this section we will discuss the recursive combination tests proposed by Brannath, Posch, and Bauer (2002). Assume p_1 and p_2 are independent and uniformly distributed random variables on [0, 1] or p-clud.

For two stage-designs, if $p_1 < \alpha_1$, reject the null; if $p_1 \geq \beta_1$, accept the null; if $\alpha_1 < p_1 \leq \beta_1$ the trial continues to the second stage. At second stage if $C\,(p_1, p_2) \leq c$, reject the null; otherwise accept the null. The combination $C\,(p_1, p_2)$ can be many different forms. To exclude nonstochastic curtailing, we assume $c < \alpha_1$; otherwise, for $\alpha_1 < p_1 \leq c$, the null hypothesis could be rejected without a second sample, although no formal stopping condition applies.

The level α test requires:

$$\alpha_1 + \int_{\alpha_1}^{\beta_1} \int_0^1 \mathbf{1}_{[C(x,y)\leq c]} dydx = \alpha, \tag{8.66}$$

where $\mathbf{1}_{[C(x,y)\leq c]}$ equals 1 if $C(x,y) \leq c$ and 0 otherwise.

For multiple-stage adaptive design, Brannath, Posch, and Bauer (2002) define a p-value for the combination test by

$$q(p_1,p_2) \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \beta_1, \\ \alpha_1 + \int_{\alpha_1}^{\beta_1} \int_0^1 \mathbf{1}_{[C(x,y)\leq c]} dydx & \text{otherwise.} \end{cases} \tag{8.67}$$

For Fisher combination, $C(p_1,p_2) = p_1 p_2$, the level α test requirement (8.66) becomes

$$c = \frac{\alpha - \alpha_1}{\ln \beta_1 - \ln \alpha_1}. \tag{8.68}$$

Carry out the integration in (8.67), we obtain

$$q(p_1,p_2) \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \beta_1, \\ \alpha_1 + p_1 p_2 (\ln \beta_1 - \ln \alpha_1) & \text{if } p_1 \in (\alpha_1, \beta_1] \text{ and } p_1 p_2 \leq \alpha_1, \\ p_1 p_2 + p_1 p_2 [\ln \beta_1 - \ln (p_1 p_2)] & \text{if } p_1 \in (\alpha_1, \beta_1] \text{ and } p_1 p_2 > \alpha_1. \end{cases} \tag{8.69}$$

The trial design is extended two-stage by two-stage until we don't want to extend any more. Let's denote the stopping boundaries by α_{i1}, β_{i1} and c_i for the i^{th} two-stage design, and the combination test p-value for the i^{th} two-stage design as $q_i(p_i, p_{i+1})$ for $i = 1, ..., K-1$, where K is the final stage, then the overall p-value can be obtained through recursion:

$$p = q_1 (p_1, q_2 (p_2, q_3 (...q_{K-1} (p_{K-1}, q_K)))) . \tag{8.70}$$

Brannath, Posch, and Bauer stated the decision roles as: we reject the null H_o, if and only if $p \leq \alpha$.

(Question for readers: Assume $K = 5$, how can we make a decision based on (8.70) at the third stage before we know $K = 5$? (8.66) implies that stopping will be based on the boundaries (α_1 and β_1), is it consistent with the stopping rule based on (8.70) if we know $K = 5$?)

The critical value (8.47) is expressed for the i^{th} two-stage design as

$$c_i = \frac{c_{i-1} - \alpha_{i1}}{\ln \beta_{i1} - \ln \alpha_{i1}}, \tag{8.71}$$

where $c_0 = \alpha$.

To plan the next two-stage design, the conditional error has to be calculated by

$$A_{i+1}(p_i) = \frac{c_i}{p_i}, \tag{8.72}$$

where $A_1(\cdot) = \alpha$.

Example 8.2 Recursive Combination Method

Brannath et el. (2002) illustrate the recursive combination test method for the hypothesis H_o:$\mu \leq 0$ against $H_a : \mu > 0$, where μ is the mean of a normally distributed random variable with known variance $\sigma^2 = 1$. For the classic design, a sample-size 144 will provide 85% power to detect an effect size of 0.25 at a one-sided level of 0.025. To use the recursive approach, the first interim is planned based on 48 patients with efficacy boundary $\alpha_{11} = 0.0102$ and futility boundary $\beta_{11} = 0.5$. This leads to the critical value $c_1 = 0.0038$ from (8.71).

Assume we get the stagewise p-value $p_1 = p_{11} = 0.199$ for the first stage, which satisfies $\beta_{11} > p_1 > \alpha_{11}$, thus the trial continues. The value of the conditional error function $A_2(p_1) = \frac{c_1}{p_1} = 0.0191$. The sample-size required for the second stage to achieve a conditional power of 0.8 is as large as 136. We decide to perform a second interim analysis after the next 68 sample units. Because the estimated mean, $\hat{\delta}_1 = \Phi^{-1}(1 - p_1)/\sqrt{n_1} = 0.122$, is not too promising, the option of stopping for futility is also taken for the second interim analysis with $a_{21} = 0.00955$ and $\beta_{21} = 0.5$. The critical value $c_2 = 0.00241$ from (8.71). Suppose we obtain the stagewise value $p_2 = p_{21} = 0.0284$ and the trial continues based on the stopping boundaries α_{21} and β_{21}. The mean estimated from the pooled samples 1 and 2 is 0.186. The conditional error is $A_3(p_2) = \frac{c_2}{p_2} = 0.0850$. The sample-size $n_3 = 68$ will provide the conditional power for the third stage 0.075 given an effect size of 0.25. We now decide not to extend any further. Assume we now observe the stagewise p-value $p_3 = 0.00781$. The overall p-value can be calculated using (8.69) as follows:

$$q_2(p_2, p_3) = 0.00955 + 0.0284\,(0.00781) \ln \frac{0.5}{0.00955} = 0.0104.$$

$$p = q_1(p_1, q_2) = 0.0102 + 0.199(0.0104)\ln\frac{0.5}{0.0102} = 0.0183.$$

Because the overall p-value $p < \alpha = 0.025$, we can reject the null hypothesis.

$$p_i(\delta_0) = 1 - \Phi\left(\left(\bar{X}_i - \delta_0\right)\sqrt{n_i}\right). \tag{8.72}$$

The computation of the confidence bounds can be pursued as follows:

(1) Based on observed data, calculate sample mean $\bar{X}_i$ for the i^{th} subsample.

(2) Select a value for δ_0 to calculate stagewise p_i using (8.72).

(3) Calculate overall p-value using (8.70).

(4) Go back to step 2 using a different δ_o until the overall p-value equals α. This δ_0 is the CI bound.

For Example 8.2, Brannath et al. calculate the 97.5% CI to be 0.0135, and the 95% CI bound 0.0446. The 50% CI is 0.1835 and gives a median unbiased point estimate for the effect size. As a comparison, the average of all the observations obtained up to the third stage equals 0.2255.

Note that when adjusting δ_0, the rejection of H_o could occur earlier than it was actually rejected. Brannath et al. (2002) didn't discuss this situation. For more details on estimations after adaptive designs, see, e.g., Posch, et al., 2005.

8.6 Decision Function Method

Müller and Schäfer (2004) generalize their early work (Müller and Schäfer, 2001; Chapter 7) to arbitrary decision function. Their early method is based on conditional rejection probability (CRP) or more precisely based on the conditional error principle, i.e., we can redesign the trial at any interim analysis as long as we keep the conditional error unchanged. Their derivation is based on the discretized Brownian motion and includes the conventional group sequential designs in Brownian motion model at the prespecified time points of the interim analysis as a special case.

We now discuss the generalization of the method. Let $X = (X_1, ..., X_k)$ denote the data collected during the experiment. Define a *decision function* $\varphi(X) = \varphi(X_1, ..., X_k) \in [0, 1]$, which can be a test statistic. At the end of the experiment the null-hypothesis H_o will be maintained if $\varphi(X) = 0$ or rejected if $\varphi(X) = 1$ is realized. For other values of $\varphi(X)$, the decision is based on a random experiment, i.e., throwing a biased coin with rejection probability given by the realized value of $\varphi(X)$. Müller

and Schäfer realize that in clinical trials that the decision based on the result of throwing a coin is problematic and suggest in the final analysis the non-randomized modification of the decision function, in which H_o will be accepted if $\varphi(X) < 1$ is realized.

We partition the X into two parts at the time of the j^{th} interim analysis: the observed dataset X_L and the planned future data X_U, $X_L \cup X_U = X$ and $X_L \cap X_U = \phi$. Define the conditional expectation of decision function as

$$\varepsilon_\theta(x_L) = E_\theta\{\varphi(X_L, X_U)\,|X_L = x_L\} \text{ for } \forall\theta \in H_o. \tag{8.73}$$

The generalized method can be simply described as follows: At any interim analysis, the trial can be redesigned with any decision function at an interim analysis as long as we keep the expected conditional decision function unchanged. The redesign procedure can be recursively applied.

8.7 Summary and Discussion

Based on the conditional error principle and closed form solutions for two-stage design, the recursive two-stage adaptive design (RTAD) approach provides an integrated process of design, monitoring, and analysis. RTAD allows trials to be designed stage by stage. At each stage, the conditional power is calculated, which is typically for monitoring, and further design can be based on this conditional power. During the redesign, information within and outside of this trial can be used. To select an optimal design for the remainder of the study at interim analysis, one can use the conditional power or utility functions. The method is applicable for general K-stage adaptive designs that allow for a very broad range of adaptations. The stopping boundary determination and the adjusted p-value calculation are straight forward with the closed forms, and no software is required.

One should be aware that every adaptive design method requires prespecification of certain aspect(s). The classic group sequential method specifies the number and timing of the analyses. The error-spending method allows for changes in the number and timing of the analyses but requires prespecification of the error-spending function. The conditional error function method prespecifies the conditional error function. The RTAD requires the prespecification that the conditional error rate will be retained at each stage based on the observed data. It is important to know that the flexibility of this method does not mean that you can design the first interim analysis arbitrarily. In fact, careful thinking and planning are required for

the initial design. Spending too much alpha at the first stage means that one has less alpha to spend at later stages, which could limit the design's ability to have good operating characteristics. Also, changing the hypothesis during a trial should be done with great caution, because it could lead to a very different clinical implication.

RTAD can also be useful in many situations. For example, it is usually unrealistic to plan many interim analyses at design stage to deal with all possible scenarios. Therefore, there are opportunities (e.g., safety concern, unexpected slow enrollment) for adding new interim analyses, changing the total number and the timing of the coming analyses after a trial is initiated, and adjusting sample-size. In these situations, RTAD is very suitable.

In addition to RTAD, we also, briefly introduce two other recursive approaches, but the computations for those methods are not as simple as RTAD.

Problem

8.2 Design trials as presented in Examples 6.1, 6.2, and 6.3 using recursive two-stage adaptive design.

8.2 Read the following publications on confidence intervals for group sequential design.

The conditional confidence interval is often very wide and inconsistent with the hypothesis testing, especially when early stopping occurs and the test statistic does not exceed the boundary with a big margin (Fan and DeMets, 2006).

The Fan-DeMets restricted conditional confidence interval (RCCI, Fan and DeMets, 2006) is narrower than ECCI but does not have the exact coverage. The unconditional exact confidence intervals are all constructed by reversing unconditional exact tests (Fan and DeMets, 2006). In general, the exact confidence interval (ECI) can be defined as

$$ECI\left(\eta = n, S_\eta = s\right) = \left\{\delta : \alpha/2 < \Pr\left(g\left(\eta, S_\eta\right) \geq g\left(n, s\right)\right) < 1 - \alpha/2\right\},$$

where the link function is presented in Table 8.2.

There are three methods of sample space ordering: (1) Stagewise ordering by Siegmund (1978), (2) Sample mean ordering by Emerson (1988) and Fleming (1990), and (3) Likelihood ratio ordering by Rosner and Tsiatis (1988) and Chang (1989). See also Strickland-Casella (2003) for the exact conditional confidence interval.

Table 8.2: Type and Link Function of Exact Confidence Intervals (ECI)

ECI type	Ordering method	Link function g
Siegmund	Stagewise ordering	$\eta + 1/\left(1 + \exp\left(-S_\eta\right)\right)$
Emerson-Fleming	Sample mean ordering	S_η/t_η
Rosner-Tsiatis-Chang	Likelihood ratio ordering	$\sqrt{t_\eta}\left(S_\eta/t_\eta - \delta\right)$
Sources:	Fan and DeMets (2006)	

Chapter 9

Sample-Size Re-Estimation Design

9.1 Opportunity

Despite a great effort, we often face a high degree of uncertainty about parameters when designing a trial or justifying the sample-size at the design stage. This could involve the initial estimates of within- or between-patient variation, a control group event rate for a binary outcome, the treatment effect desired to be detected, the recruiting pattern, or patient compliance, all of which affect the ability of the trial to address its primary objective (Shih, 2001). This uncertainty can also include the correlation between the measures (if a repeated measure model is used) or among different variables (multiple endpoints, covariates). If a small uncertainty of prior information exists, a classic design can be applied. However, when the uncertainty is greater, a classic design with a fixed sample-size is inappropriate. Instead, it is desirable to have a trial design that allows for re-estimation of sample-size in the middle of the trial based on unblinded data. Several different algorithms have been proposed for sample-size re-estimation, including the conditional power approach and Cui-Hung-Wang's approach based on the ratio of observed effect size versus the expected effect size.

In this chapter, we will evaluate the performance of different sample-size modification methods. Operationally, it is a concern that sample-size re-estimation will release the unblinded efficacy data to the general public prematurely. Using a discrete function for sample-size re-estimation is suggested, such that the exact effect size would not be revealed. We will study the impact on efficiency of this information-mask approach. The adjusted p-value, the point estimate and confidence interval calculation will also be discussed.

It is important to differentiate the two different properties: those properties at the design stage and those at the interim analyses. For example, power is an interesting property at the design stage, but at the time of in-

terim analysis, power has little value, and the conditional power is of great concern. From a statistical point of view, most adaptive design methods do not require a prespecification of sample-size adjustment rules at design stage. How to adjust the sample size can be determined right after the interim analysis.

In later of the chapter, we'll give two examples using SSR: a myocardial infarction prevention trial and a non-inferior adaptive trial with Farrington-Manning margin. Summaries and discussions will be presented in the last section.

9.2 Adaptation Rules

There are many possible rules for sample-size adjustment. Here we will discuss only two types of adjustments: (1) sample-size adjustment based on the effect-size ratio between the initial estimate and the observed estimate, and (2) sample-size adjustment based on conditional power.

9.2.1 *Adjustment Based on Effect Size Ratio*

The formation for sample-size adjustment based on the ratio of the initial estimate of effect size (E_0) to the observed effect size (E) is given by

$$N = \left| \frac{E_0}{E} \right|^a N_0, \tag{9.1}$$

where N is the newly estimated sample-size per group, N_0 is the initial sample-size per group which can be estimated from a classic design, and $a > 0$ is a constant,

$$E = \frac{\hat{\eta}_{i2} - \hat{\eta}_{i1}}{\hat{\sigma}_i}. \tag{9.2}$$

With a large sample assumption, the common variance for the two treatment groups is given by (See Chapter 2)

$$\hat{\sigma}_i^2 = \begin{cases} \hat{\sigma}_i^2 & \text{for normal endpoint,} \\ \bar{\eta}_i(1-\bar{\eta}_i) & \text{for binary endpoint,} \\ \bar{\eta}_i^2 \left[1 - \frac{e^{\bar{\eta}_i T_0} - 1}{T_0 \bar{\eta}_i e^{\bar{\eta}_i T_s}}\right]^{-1} & \text{for survival endpoint,} \end{cases} \tag{9.3}$$

where $\bar{\eta}_i = \frac{\hat{\eta}_{i1} + \hat{\eta}_{i2}}{2}$ and the logrank test is assumed to be used for the survival analysis. Note that the standard deviations for proportion and

survival have several versions. There are usually slight differences in the resulting sample-size or power among the different versions.

The sample-size adjustment in (9.1) should have additional constraints: (1) It should be smaller than $N_{\max}$ (due to financial and or other constraints) and greater than or equal to $N_{\min}$ (the sample-size for the interim analysis) and (2) If E and E_0 have different signs at the interim analysis, no adjustment will be made.

To avoid numerical overflow when $E = 0$, the actual algorithm implemented in SAS macros in this chapter is

$$N = min\left\{N_{max},\, max\left(N_0, \frac{E_0}{abs(E) + 0.0000001} N_0\right)\right\}. \tag{9.4}$$

9.2.2 *Adjustment Based on Conditional Power*

For an SSR design, the conditional power is more important than the power. The conditional power for a two-stage design is given by (Chapter 7)

$$cP = 1 - \Phi\left(B\left(\alpha_2, p_1\right) - \tfrac{\delta}{\sigma}\sqrt{\tfrac{n_2}{2}}\right),\ \alpha_1 < p_1 \le \beta_1.$$

The formulation for the new sample-size for a two stage design is discussed in Chapter 7 and given by

$$n_2 = \frac{2\sigma^2}{\delta^2}\left[B\left(\alpha_2, p_1\right) - \Phi^{-1}\left(1 - cP\right)\right]^2, \tag{9.5}$$

where cP is the target conditional power and function $B\left(\alpha_2, p_1\right)$ is given in Table 9.1 for different design methods. The actual algorithm implemented in SAS Macro 9.1 is:

$$n_2 = \begin{cases} max\left\{N_{2\,\min}, \frac{2\sigma^2}{\delta^2}\left[B\left(\alpha_2, p_1\right) - \Phi^{-1}\left(1 - cP\right)\right]^2\right\} & \text{If } \alpha_1 < p_1 < \beta_1 \\ 0, & \text{otherwise,} \end{cases} \tag{9.6}$$

where $\delta = \left(\mu_2 - \mu_1\right)/\sigma$, n_2 is the sample-size per group at stage 2 with a minimum of $N_{2\,\min}$. Assume a known σ.

For a K-stage design, (9.5) is an approximation. Suppose we are interested only in the case where SSR takes place only at the first interim analysis, and the sample-size change affects only the last stage sample-size.

Note that (9.4) allows only a sample-size increase from N_0, but (9.6) allows both an increase and decrease in sample-size. The futility boundary

is suggested because (1) a certain small δ_1 will virtually never lead to statistical significance at the final analysis, and (2) a small $\hat{\delta}$ at the final analysis will be clinically unjustifiable; therefore, adding a futility boundary could be cost-saving.

Table 9.1: Function $B(\alpha_2, p_1)$ for Conditional Power

Design Method	$B(\alpha_2, p_1)$
MIP	$\Phi^{-1}(1-\alpha_2)$
MSP	$\Phi^{-1}(1-\max(0, \alpha_2 - p_1))$
MLP	$\Phi^{-1}\left(1-\max(0, \frac{\alpha_2}{w_2} - \frac{w_1 p_1}{w_2})\right)$
MPP	$\Phi^{-1}\left(1-\frac{\alpha_2}{p_1}\right)$
MINP	$\frac{\sqrt{w_1^2+w_2^2}\Phi^{-1}(1-\alpha_2)-w_1\Phi^{-1}(1-p_1)}{w_2}$

9.3 SAS Macros for Sample-Size Re-estimation

SAS Macro 9.1 is developed to simulate a two-arm, K-stage adaptive design with a normal, binary, or survival endpoint using "MIP", "MSP", "MPP", or "LW". The sample-size adjustment is based on the conditional power method (9.6) for a two-stage design or (9.4) for K-stage design (K>2) . The sample-size adjustment is allowed only at the first interim analysis, and the sample-size adjustment affects only the final stagewise sample-size. **ux** and **uy** = the means, response rates, or hazard rates for the two groups, and **Ns{k}** = sample-size for group x at stage k. The random increment in sample-size of **nMinIcr** is added for the information mask, **nMinIcr** = minimum sample-size increment in group x for the conditional power approach only, **n2new** = the re-estimated sample-size for group x at the second stage, and **eSize** = the standardized effect size. **nSims** = number of simulation runs, **nStgs** = number of stages, **alpha0** = overall α, and **EP** = "normal", "binary", or "survival". **Model** = "MIP", "MSP", "MPP", or "LW"; **Nadj** = "N" for the case without SSR; **Nadj** = "Y" for the case with SSR. **cPower** = target conditional power, **DuHa** = estimated treatment difference, **Nmax** = the maximum sample-size allowed for group x, **N2min** = minimum sample size (not cumulative) in group x for stage 2, **sigma** = standard deviation for normal endpoint, **tAcr** = accrual time, **tStd** = study duration, and **power** = initial target power for the trial, **nRatio** = the sample size ratio of group y to group x for an imbalanced design. **NIType** = "FIXED" for non-inferiority design with a fixed NI margin and **NIType** = "PCT" for non-inferiority design with Farrington-Manning NI margin, **NId** = NI margin. **Aveux**, **Aveuy**, and **AveN** =

average simulated responses (mean, proportion, or hazard rate) and sample size, **FSP{i}** = futility stopping probability at the i^{th} stage, **ESP{i}** = efficacy stopping probability at the i^{th} stage, and **alpha{i}** and **beta{i}** = efficacy and futility stopping boundaries at the i^{th} stage.

››SAS Macro 9.1: Adaptive Design with Sample-Size Re-estimation››

```
%Macro nByCPowerUB(Model, a2, eSize, cPower, p1, w1, w2, n2New);
 If &Model="MIP" Then BFun=Probit(1-&a2);
  If &Model="MSP" Then BFun=Probit(1-max(0.0000001,&a2-&p1));
 If &Model="MPP" Then BFun=Probit(1-&a2/&p1);
  If &Model="LW" Then
      BFun=(Probit(1-&a2)- &w1*Probit(1-&p1))/&w2;
 &n2New=2*((BFun-Probit(1-&cPower))/&eSize)**2; *n for group x;
%Mend nByCPowerUB;

%Macro TwoArmNStgAdpDsg(nSims=1000000, nStgs=2,ux=0, uy=1,
 NId=0, NItype="FIXED", EP="normal", Model="MSP",
 Nadj="Y", cPower=0.9, DuHa=1, Nmax=300, N2min = 150,
 nMinIcr=1, sigma=3, tAcr=10, tStd=24, nRatio=1);
DATA NStgAdpDsg; Set dInput;
KEEP Model power Aveux Aveuy AveTotalN FSP1-FSP&nStgs
     ESP1-ESP&nStgs;
Array Ns{&nStgs}; Array alpha{&nStgs}; Array beta{&nStgs};
Array ESP{&nStgs}; Array FSP{&nStgs}; Array Ws{&nStgs};
Array sumWs{&nStgs}; Array TSc{&nStgs};
Model=&Model; cPower=&cPower; nRatio=&nRatio; NId=&NId;
NItype=&NItype; N2min=&N2min; nStgs=&nStgs; sigma=&sigma;
power=0; AveTotalN=0; Aveux=0; Aveuy=0; ux=&ux; uy=&uy;
cumN=0; Do i=1 To nStgs-1; cumN=cumN+Ns{i}; End;
N0=CumN+Ns{nStgs};
Do k=1 To nStgs;
   sumWs{k}=0; Do i=1 To k; sumWs{k}=sumWs{k}+Ws{i}**2; End;
   sumWs{k}=sqrt(sumWs{k});
End;
* Calcate the standard deviation, sigma for different endpoints *;
 If &EP="normal" Then Do sigmax=&sigma; sigmay=&sigma; End;
 If &EP="binary" Then Do
     sigmax=Sqrt(&ux*(1-&ux)); sigmay=Sqrt(&uy*(1-&uy));
 End;
 If &EP="survival" Then Do
```

```
    sigmax=ux*sqrt(1+exp(-ux*&tStd)*(1-exp(ux*&tAcr))/(&tAcr*ux));
    sigmay=uy*sqrt(1+exp(-uy*&tStd)*(1-exp(uy*&tAcr))/(&tAcr*uy));
  End;
  If NItype="PCT" Then sigmax=(1-NId)*sigmax;

 Do i=1 To nStgs; FSP{i}=0; ESP{i}=0; End;
 Do iSim=1 to &nSims;
  ThisNx=0; ThisNy=0; Thisux=0; Thisuy=0;
  Do i=1 To nStgs; TSc{i}=0; End;
  TS=0; If &Model="MPP" Then TS=1;

  Do i=1 To nStgs;
    uxObs=Rannor(746)*sigmax/sqrt(Ns{i})+&ux;
    uyObs=Rannor(874)*sigmay/sqrt(nRatio*Ns{i})+&uy;

    Thisux=Thisux+uxObs*Ns{i};
    Thisuy=Thisuy+uyObs*nRatio*Ns{i};
    If NItype="PCT" Then NId=uxObs*&NId;
    ThisNx=ThisNx+Ns{i};
    ThisNy=ThisNy+Ns{i}*nRatio;
    StdErr=(sigmax**2/Ns{i}+sigmay**2/(nRatio*Ns{i}))**0.5;
    TS0 = (uyObs-uxObs+NId)/StdErr;

    If Model="MIP" Then TS=1-ProbNorm(TS0);
    If Model="MSP" Then TS=TS+(1-ProbNorm(TS0));
    If Model="MPP" Then TS=TS*(1-ProbNorm(TS0));
    If Model="LW" Then Do;
      Do k=i to nStgs; TSc{k}=TSc{k}+Ws{i}/sumWs{k}*TS0; End;
      TS=1-ProbNorm(TSc{i});
    End;
    If Model="UWZ" Then Do;
      StdErr=(sigmax**2/ThisNx+sigmay**2/ThisNy)**0.5;
      TS0=(Thisuy/ThisNy-Thisux/ThisNx+NId)/StdErr;
      TS=1-ProbNorm(TS0);
    End;
    If TS>beta{i} Then Do; FSP{i}=FSP{i}+1/&nSims; Goto Jump; End;
    Else If TS<=alpha{i} then do;
      Power=Power+1/&nSims; ESP{i}=ESP{i}+1/&nSims;
          Goto Jump; End;
    Else If nStgs>1 & i=1 & &Nadj="Y" Then Do;
      eSize=&DuHa/(abs(uyObs-uxObs)+0.0000001);
```

```
    nFinal=min(&Nmax, max(N0,eSize*Abs(eSize)*N0));

      If nStgs=2 Then do;
      eSize=(uyObs-uxObs+NId)/((sigmax+sigmay)*sqrt(0.5+0.5/nRatio));
      %nByCPowerUB(Model, alpha{2}, eSize, cPower, TS,
            ws{1}, ws{2}, n2New);
      nFinal=Round(min(&Nmax,ns{1}+n2New+&nMinIcr/2), &nMinIcr);
      nFinal=max(N2min+Ns{1},nFinal);
      End;

      If nStgs>1 Then Ns{nStgs}= max(1,nFinal-cumN);
    End;
  End;

Jump:
Aveux=Aveux+Thisux/ThisNx/&nSims;
Aveuy=Aveuy+Thisuy/ThisNy/&nSims;
AveTotalN=AveTotalN+(ThisNx+ThisNy)/&nSims;
End;
Output;
Run;
Proc Print; Run;
%Mend TwoArmNStgAdpDsg;
```

‹‹SAS‹‹

Note that SAS Macro 9.1 actually includes two macros: nByCPowerUB and TwoArmNStgAdpDsg; the latter calls the former.

9.4 Comparison of Sample-Size Re-estimation Methods

We now use SAS Macro 9.1 to study the operating characteristics of various sample-size re-estimation methods for two-stage adaptive designs. The non-binding futility rule is currently adopted by the regulatory bodies, based on which the futility boundaries don't have to be followed. Therefore, the earlier futility boundaries cannot be considered in constructing later stopping boundaries. For this reason, it is important to study the performance of different methods with non-binding futility boundaries. We will compare the power and sample size among different methods. The scenarios considered are: (1) the trial is properly powered; (2) the trial is under-powered; (3) the trial has a futility boundary, and (4) the sample-size increases dis-

cretely for information mask. For each of the scenarios, we simulate the classic and adaptive designs using the following three different methods: MSP, MPP, and MINP. The simulations are performed for a Normal endpoint with a mean $u_A = 0$ for the control group, $u_B = 1$ for the test group, and a common standard deviation of sigma = 3.

The results in Tables 9.2 are generated using the following three SAS macro calls. The SAS macro calls for generating results in Tables 9.3 through 9.5 are similar and are not presented. $\bar{N}$ is the average total sample-size from the simulations, and $\bar{u}_A$ and $\bar{u}_B$ are the average means in group A and group B from the simulations. Each scenario has 1,000,000 simulation runs. ESP1 is the early stopping probability at the first stage.

```
>>SAS>>
Title "High Initial Power with MSP, Ha - Table 9.2";
Data dInput;
Array Ns{2} (100,100); Array alpha{2} (0.0025,0.21463);
Array beta{2} (0.2146,0.2146); Array Ws{2} (1,1);
%TwoArmNStgAdpDsg(ux=0, uy=1, N2min=100, EP="normal",
  Model="MSP", Nadj="Y", cPower=0.9, DuHa=1, sigma=3,
  Nmax=300, nMinIcr=20);
Run;

Title "High Initial Power with MPP, Ha - Table 9.2";
Data dInput;
Array Ns{2} (100,100); Array alpha{2} (0.0025,0.00375);
Array beta{2} (0.2146,0.00375); Array Ws{2} (1,1);
%TwoArmNStgAdpDsg(ux=0, uy=1, N2min=100, EP="normal",
  Model="MPP", Nadj="Y", cPower=0.9, DuHa=1, sigma=3,
  Nmax=300, nMinIcr=20);
Run;

Title "High Initial Power with MINP, Ha - Table 9.2";
Data dInput;
Array Ns{2} (100,100); Array alpha{2} (0.0025,0.024);
Array beta{2} (0.2146,0.024); Array Ws{2} (1,1);
  %TwoArmNStgAdpDsg(ux=0, uy=1, N2min=100, EP="normal",
    Model="LW",  Nadj="Y", cPower=0.9, DuHa=1, sigma=3,
    Nmax=300, nMinIcr=20);
Run;
<<SAS<<
```

Scenario 1: High Initial Power

A classic design with 400 subjects will provide 91.5% power to detect an effect size of 1/3 at a one-sided level $\alpha = 0.025$. For the two-stage SSR designs with a maximum of 600 subjects, the minimum sample-size for the second stage, N2min = 100/group, and $\alpha_1 = 0.0025$. The simulation results are presented in Tables 9.2.

Table 9.2: Comparison of Simulation Results Under Ha

Method	α_2	ESP1	FSP1	Power	$\bar{N}$	$\bar{u}_A$	$\bar{u}_B$
MSP	.21463	.326	.059	.933	446	-0.03	1.03
MPP	.00375	.326	.059	.927	424	-0.03	1.03
LW	.02400	.326	0.059	.926	388	-0.03	1.03

Note: Initial sample-size per group: n1=n2=200, nMinIcr=20.

From Table 9.2, we can see that the power and the expected sample size are larger than those in the classic design. The bias of the naive is about 6%.

Scenario 2: Low Initial Power

We now study the characteristics of the adaptive designs with a low initial power. In this scenario, the true mean difference is 1 with a standard deviation of 3, but the treatment difference is over estimated as 1.25. A classic design with 240 subjects per group will have 73.5% power to detect an effect size of 1/3 at a one-sided level $\alpha = 0.025$. For the two-stage adaptive designs with maximum 600 subjects (N2min = 60) simulation results are presented in Table 9.3.

From Table 9.3, we can see that the power is protected (>80%) by all methods. The bias of the naive is negligible.

Table 9.3: Comparison of Simulation Results Under H_a

Method	α_2	ESP1	FSP1	Power	$\bar{N}$	$\bar{u}_A$	$\bar{u}_B$
MSP	.21463	.163	.150	.843	428	-0.00	1.00
MPP	.00375	.163	.150	.838	404	-0.01	1.01
LW	.02400	.163	.150	.823	336	-0.01	1.01

Initial sample-size per group: n1=n2=120, nMinIcr=20.

Scenario 3: Effect of Early Stopping Boundary

Let's use $\alpha_1 = 0.005$, $\beta_1 = \alpha_2 = 0.205$, and $N_1 = 105$ for MSP, and $\alpha_1 = 0.005$, $\beta_1 = 1$, $N_1 = 60$ for MPP and LW. Note that when $p_1 > \beta_1 = 0.205$, there is no possibility of rejecting the null hypothesis at the second stage for MSP. We choose a larger sample-size N_1 at the first stage for MSP because we want to have similar power for all methods. The maximum sample-size

$N_{\max}$ is 200 per group. Table 9.4 gives a quick comparison of different methods, where $\bar{\text{N}}$o is the expected sample-size under the null and $\bar{\text{N}}$a is the expected sample-size under the alternative. All designs have a 90% target conditional power. Overall, MSP, MPP, and LW perform equally well under the alternative hypothesis. However, MSP has much smaller sample-size than others under the null.

Table 9.4: Comparisons of Adaptive Methods

Method	α_2	$\bar{\text{N}}$o	$\bar{\text{N}}$a	Power
MSP	.2050	248	306	88.9%
MPP	.0038	398	323	89.4%
LW	.0226	398	323	89.6%

Note: α_1= 0.005, effect size =1/3.

Scenario 4: Discrete SSR for Information-Mask

To study the impact of the discrete SSR, we study the minimum sample-size increment from 1, 20 to 50 per group. The simulation results are presented in Table 9.5. We can see that the discretization of sample-size increment has minimal impact on the power and sample-size for the adaptive designs.

Table 9.5: Summary of Comparisons with Lower Initial Power

	$\Delta N_{\min}$= 1		$\Delta N_{\min}$= 20		$\Delta N_{\min}$= 50	
Method	$\bar{\text{N}}$	Power	N	Power	N	Power
MSP	424	0.842	428	0.843	435	0.843
MPP	401	0.838	404	0.838	408	0.839
LW	331	0.820	336	0.823	344	0.827

Note: α_1= 0.05 and β_1= 0.25, N0 = 120, Nmax = 300.

Note that comparison based on power is not the best way for adaptive designs. The power depends on many different things: (1) the unknown true treatment difference δ, (2) the sample size at stage 1, and (3) the rule for SSR. Because δ is an unknown, we may use Bayesian prior distribution $\pi(\delta)$ (Chapter 16). Note that there are two types of conditional power we have considered: (1) conditioning on p_1 and (2) conditioning on p_1 and $\delta = \hat{\delta}_1$. The former is what we are really interested, while the latter is just an estimation of the former. Just like power, there are two types of power: (1) the power depending on true δ and (2) the power depending on $\hat{\delta}$ that is estimated at the initial design. Therefore, we should not make general conclusions regarding which method is better just based on the power.

Scenario 5: Comparison of Conditional Power

As stated earlier, for adaptive design, conditional power is a better measure than power regarding the efficiency. The difference in conditional power between different methods is dependent on the stagewise p-value from the first stage. From Tables 9.6 and 9.7 and Figure 9.1, it can be seen that conditional power for MSP is uniformly higher for p_1 around 0.1 than the other two methods. Therefore, if you believe that p_1 is somewhere between (0.005, 0.18), then MSP is much efficient than MPP and MINP or LW; otherwise, MPP and LW are better.

Table 9.6: Conditional Power as Function of N_2

		$N_2 = 100$	$N_2 = 200$	$N_2 = 300$	$N_2 = 400$
Method	α_2	Power	Power	Power	Power
MSP	.21463	0.584	0.788	0.894	0.948
MPP	.00375	0.357	0.567	0.748	0.853
LW	.02400	0.460	0.686	0.825	0.906

Note: $\alpha_1 = 0.0025$. $p_1 = 0.1$, effect size = 0.2, no futility binding.

Examples of SAS macro calls for generating the results in Tables 9.6 are presented in the following:

```
>>SAS>>
%ConPower(EP="normal", Model="MSP", alpha2=.21463,
   ux=0.2, uy=0.4, sigma=1, n2=200, p1=0.1);
%ConPower(EP="normal", Model="MPP", alpha2=.00375,
   ux=0.2, uy=0.4, sigma=1, n2=200, p1=0.1);
%ConPower(EP="normal", Model="LW", alpha2=.02400,
   ux=0.2, uy=0.4, sigma=1, n2=200, p1=0.1, w1=1, w2=1);
<<SAS<<
```

Table 9.7: Conditional Powers as Function of P_1

				p_1		
Method	0.010	0.050	0.100	0.015	0.018	0.220
				Power		
MSP	0.880	0.847	0.788	0.685	0.572	0.000
MPP	0.954	0.712	0.567	0.516	0.485	0.453
LW	0.937	0.802	0.686	0.595	0.547	0.490

Note: $\alpha_1 = 0.0025$. $N_2 = 200$, effect size = 0.2, no futility binding.

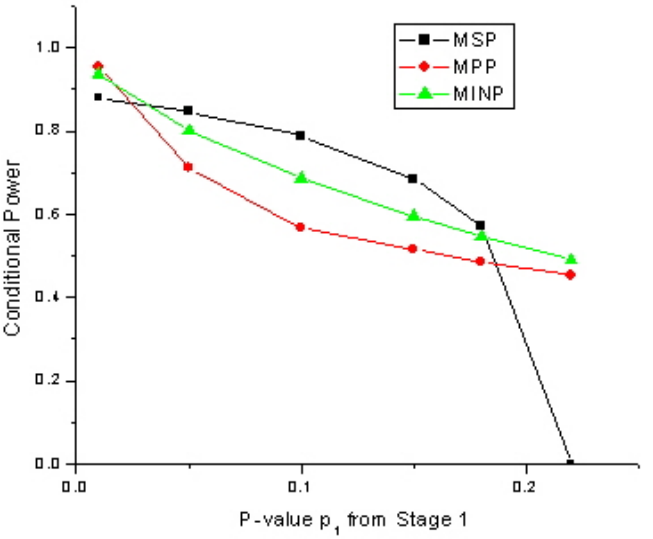


Figure 9.1: Conditional Power versus P-value from Stage 1

9.5 Analysis of Design with Sample-Size Adjustment

9.5.1 *Adjusted P-value*

We recommend the stagewise-ordering p-value that has been discussed in Chapters 4 and 5. For MSP, we use (4.9), i.e.,

$$p(t;k) = \begin{cases} t, & k = 1, \\ \alpha_1 + t(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2), & k = 2 \text{ and } \beta_1 < \alpha_2, \\ \alpha_1 + \frac{1}{2}(t - \alpha_1)^2, & k = 2 \text{ and } \beta_1 \geq \alpha_2. \end{cases} \tag{9.7}$$

For MPP, we use (4.16), i.e.,

$$p(t;k) = \begin{cases} t, & k = 1, \\ \alpha_1 + t \ln \frac{\beta_1}{\alpha_1}, & k = 2\ . \end{cases} \tag{9.8}$$

For MINP, we use simulations to determine the stagewise-ordering p-value.

9.5.2 *Confidence Interval*

From the duality of the confidence and hypothesis test, we know that a $100(1-\alpha)\%$ CI consists all δ_0 such that the null hypothesis $H_o : \delta \leq \delta_0$ is not rejected at any stage before stage K_s, where K_s is the stage at which the trial was actually stopped.

The stagewise p-value at the i^{th} stage for the one-sided null hypothesis $H_o : \delta \leq \delta_0$ is given by

$$p_i = 1 - \Phi\left(\frac{\hat{\delta}_i - \delta_0}{\sigma}\sqrt{\frac{n_i}{2}}\right), \tag{9.9}$$

where $\hat{\delta}_i$ is the naive estimate based on the i^{th} stage subsample.

Therefore, for MSP, the stagewise CI limits can be obtained by solving (9.1) for δ_{α_k}:

$$\sum_{i=1}^{k}\left[1 - \Phi\left(\frac{\hat{\delta}_i - \delta_{\alpha_k}}{\sigma}\sqrt{\frac{n_i}{2}}\right)\right] = \alpha_k,\ k = 1, ..., K_s, \tag{9.10}$$

where stage K_s is the stage where the trial was actually stopped.

For MPP, the stagewise CI limits can be obtained by solving (9.11) for δ_{α_k}:

$$\prod_{i=1}^{k}\left[1 - \Phi\left(\frac{\hat{\delta}_i - \delta_{\alpha_k}}{\sigma}\sqrt{\frac{n_i}{2}}\right)\right] = \alpha_k,\ k = 1, ..., K_s. \tag{9.11}$$

For MINP, the stagewise CI limits can be obtained by solving (9.12) for δ_{α_k}:

$$\sum_{i=1}^{k} w_{ki}\left(\frac{\hat{\delta}_i - \delta_{\alpha_k}}{\sigma}\sqrt{\frac{n_i}{2}}\right) = \alpha_k,\ k = 1, ..., K_s, \tag{9.12}$$

where w_{ki} are predetermined weights satisfying $\sum_{i=1}^{k} w_{ki}^2 = 1$.

The $100(1-\alpha)\%$ confidence limit $\delta_{0\,\min}$ is given by

$$\delta_{0\,\min} = \max\left\{\delta_{\alpha_1}, ..., \delta_{\alpha_{K_s}}\right\}. \tag{9.13}$$

(9.13) is also a stagewise-ordering $100(1 - \sum_{i=1}^{K_s}\pi_i)\%$ CI (see Chapter 3).

Note that (9.12) reduces to the classic CI when $K = 1$. (9.12) can be solved analytically. For example, for the second stage, we have

$$\delta_{\alpha_2} = \frac{w_{k1}\sqrt{n_1}\hat{\delta}_1 + w_{k2}\sqrt{n_2}\hat{\delta}_2 - \sqrt{2}\sigma z_{1-\alpha_2}}{w_{k1}\sqrt{n_1} + w_{k2}\sqrt{n_2}}. \tag{9.14}$$

So far, we have not considered the futility boundaries in constructing CI. To use futility boundaries in constructing CI, see Problem 9.7.

9.5.3 *Adjusted Point Estimates*

Design without Possible Early Stopping

When there is no early stopping, e.g., the interim analysis is for SSR only, then the unbiased estimate can be easily found (Liu and Proschan, 2002; Proschan, 2003; Brannath, Konig, and Bauer, 2006). For example, the following weighted stagewise estimate is unbiased:

$$\hat{\delta}_u = \sum_{i=1}^{K} \omega_i \hat{\delta}_i, \tag{9.15}$$

where $\hat{\delta}_i$ is the naive estimate based on the i^{th} stage subsample, e.g., mean difference or response rate difference between the two groups. Note that $E(\delta_i) = \delta$ $(i = 1, ..., K)$, hence for any predetermined constant weights ω_i satisfying $\sum_{i=1}^{K} \omega_i = 1$ will provide a unbiased estimate. However, for consistency with the CI, the weight should be carefully chosen. If the SSR trial only allows for sample-size increase such that the final sample-size is between the initial sample-size N_0 and the maximum sample-size $N_{\max}$, then the following weights might be a good choice for the two-stage SSR design:

$$\omega_1 = \frac{2n_1}{N_0 + N_{\max}},\ \omega_2 = 1 - \omega_1. \tag{9.16}$$

Design with Possible Early Stopping

If symmetric stopping boundaries are used, an unbiased point estimate is given by (3.36) in Chapter 3. For a two-stage design, we can use point estimate:

$$\hat{\delta}_u = \frac{w_1\sqrt{n_1}\hat{\delta}_1 + w_2\sqrt{n_2}\hat{\delta}_2}{w_1\sqrt{n_1} + w_2\sqrt{n_2}}, \tag{9.17}$$

where w_i are predetermined constants, which is suggested to be $w_1 = \sqrt{\frac{n_1}{n_1+n_{02}}}$ and $w_2 = \sqrt{\frac{n_{02}}{n_1+n_{02}}}$. Here n_{02} is the initial sample-size for the

second stage (Lawrence and Hung, 2003).

For a general adaptive design, there is an absolute minimum sample-size $n_{\min}$ required regardless of the adaptations. We can use the first $n_{\min}$ patients per group to construct an unbiased point estimate:

$$\hat{\delta}_u = \sum_{i=1}^{n_{\min}} \frac{x_{Bi}}{n_i} - \sum_{i=1}^{n_{\min}} \frac{x_{Ai}}{n_i}, \tag{9.18}$$

where x_{Ai} and x_{Bi} are the observed responses of the i^{th} patient in groups A and B, respectively.

Equation (9.18) is difficult to justify: Why are only the first $n_{\min}$ patients' outcomes considered? If we believe bias is not the only issue in drug development, we should balance the bias and interpretability of the results. I'd suggest that readers go back to Chapter 3 for more discussions on this. Also, there will be a more in-depth discussion in Chapter 18.

Using an estimate that has a median equal to δ independently from the adaptation (Brannath, Konig and Bauer, 2002; Lawrence and Hung, 2003; Proschan, 2003; and Cheng and Shen, 2004) has also been suggested.

For a design featuring early stopping, (9.17) is an unbiased median, but the mean bias and variance of $\hat{\delta}_u$ depend on the adaptation rule.

The maximum bias for a two-stage adaptive design with early stopping and SSR is bound by (See Problem 9.1)

$$0.4\frac{\sigma}{\sqrt{n_1}}\frac{N_{\max}-n_1}{N_{\max}} = \frac{0.4\sigma}{\sqrt{n_1}}\frac{r_{\max}-1}{r_{\max}} < \frac{0.4\sigma}{\sqrt{n_1}}, \tag{9.19}$$

where $N_{\max}$ is the maximum sample-size per group and $r_{\max} = N_{\max}/n_1$.

An interesting method to obtain an unbiased estimate is to recruit at least two more patients for each stage $i > k$ when the trial is stopped at the k^{th} stage due to efficacy or futility such that we use (9.15) and (9.16). However, the extra patients enrolled will not be used for the hypothesis test. However, when the trial stops at the first stage, there could be a consistency issue between point estimation and hypothesis test.

Another method is to use the linear bias correction approach described in Chapter 3 by means of computer simulation.

9.6 Trial Example: Prevention of Myocardial Infarction

Example 9.1 Myocardial Infarction Prevention Trial

This example is based on the case presented by Cui, Hung, and Wang

(1999). In a phase-III, two-arm trial to evaluate the effect of a new drug on the prevention of myocardial infarction in patients undergoing coronary artery bypass graft surgery, a sample-size of 300 patients per group will provide 95% power to detect a 50% reduction in incidence from 22% to 11% at the one-sided significance level $\alpha = 0.025$. Although the sponsor is confident about the incidence of 22% in the control group, the sponsor is not that sure about the 11% incidence rate in the test group.

Because of the wide range of the estimated incidence rate in the test group, the sponsor felt uncomfortable choosing a fixed sample-size. A fixed sample-size can be either too small for a small effect size or too large for the large effect size. Therefore, an adaptive design with SSR is used. Further, assume a safety requirement for a minimum number of patients to be treated that precludes interim efficacy stopping. Therefore an adaptive trial with futility stopping and SSR is chosen.

Suppose we decide to use MSP for a two-stage adaptive design featuring sample-size re-estimation. The conditional power approach (9.5) will be used for sample-size adjustment based on the interim analysis results, which is scheduled when efficacy assessments have been completed for 50% of patients. The stopping rules are: at stage 1, stop for futility if the stagewise p-value $p_1 > \beta_1$, and stop for efficacy if $p_1 \leq \alpha_1$; at the final stage, if $p_1 + p_2 \leq \alpha_2$, claim efficacy; otherwise claim futility. The designs with and without SSR will be evaluated using stopping boundaries: $\alpha_1 = 0$, $\beta_1 = 0.2$, $\alpha_2 = 0.225$. The upper limit of the sample-size is $N_{\max} = 500$ per group. The SAS macro calls for the simulations are shown as follows:

```
>>SAS>>
Title "Example 9.1";
Data dInput;
Array Ns{2} (150, 150); Array alpha{2} (0,0.225);
Array beta{2} (0.2,0.225); Array Ws{2} (1,1);
%TwoArmNStgAdpDsg(nStgs=2, ux=0.11, uy=0.22,
EP="binary", Model="MSP", Nadj="N");

%TwoArmNStgAdpDsg(nStgs=2, ux=0.11, uy=0.22,
EP="binary", Model="MSP", Nadj="Y", cPower=0.95,
DuHa=0.11, Nmax=500, N2min=150, nMinIcr=50);

%TwoArmNStgAdpDsg(nStgs=2, ux=0.14, uy=0.22,
EP="binary", Model="MSP", Nadj="N");

%TwoArmNStgAdpDsg(nStgs=2, ux=0.14, uy=0.22,
```

```
EP="binary", Model="MSP", Nadj="Y", cPower=0.95,
DuHa=0.11, Nmax=500, N2min=150, nMinIcr=50);

%TwoArmNStgAdpDsg(nStgs=2, ux=0.22, uy=0.22,
EP="binary", Model="MSP", Nadj="N");

%TwoArmNStgAdpDsg(nStgs=2, ux=0.22, uy=0.22,
EP="binary", Model="MSP", Nadj="Y", cPower=0.95,
DuHa=0.11, Nmax=500, N2min=150, nMinIcr=50);
Run;
```
‹‹**SAS**‹‹

The simulation results are presented in Table 9.8. From the simulation results, there are several noticeable features of adaptive designs:

(1) For the GSD with futility boundary $\beta_1 = 0.2$ without SSR, there is a reduction in sample-size under H_o as compared to the classic design, but a decrease in power as compared to the classic design, as well.

(2) For the designs with SSR, the power and sample-size increase when the treatment effect is overestimated.

Table 9.8: Comparison of Adaptive Designs

	Event Rate in the Test Group P_T				
Design	0.110		0.14		0.22
	$\bar{N}$a	Power (%)	$\bar{N}$a	Power (%)	$\bar{N}$o
Classic	600	94.2	600	72.3	600
GSD	588	92.0	550	67.0	360
SSR	928	95.8	874	80.0	440

Note: $\alpha_1 = 0$, $\beta_1 = 0.2$, $N_{\max} = 500$/group, target cP = 0.95.

Based on the simulation results, we suggest using the adaptive design with SSR. We promise that if the trial stops at the first stage, two more patients will be enrolled for the estimation. Therefore we can use (9.15) and (9.16).

Regardless of whether the trial is stopped at the first stage or the second stage, we have two stagewise naive estimates, $\hat{\delta}_1$ and $\hat{\delta}_2$. Suppose an interim analysis is performed with 150 patients per group. The observed event rates are 0.22 and 0.165 for the control and test groups, respectively. The p-value can be calculated using the Chisq test or equivalently $p_1 = 1 - \Phi\left(\frac{(0.22-0.165)\sqrt{150}}{\sqrt{0.31}}\right) = 1 - \Phi(1.2111) = 0.1129 < \beta_1 = 0.2$. Therefore, the trial proceeds to the second stage with a newly estimated sample-size per group for the second stage $n_2 = N_{\max} - n_1 = 650$ for a target

conditional power of 90% with effect size of 0.14. This sample size is considered too big due to financial consideration. $n_2 = 400$ is finally used, which provides about 78% conditional power. Suppose the observed event rates for Stage 2 are 0.22 and 0.175 for the control and test groups, respectively; the stagewise p-value is $p_2 = 1 - \Phi\left(\frac{(0.22-0.175)\sqrt{400}}{\sqrt{0.316}}\right) = 1 - \Phi(1.6011) = 0.0547$. Therefore, the test statistic $t = p_1 + p_2 = 0.1676 < \alpha_2 = 0.225$ and the null hypothesis is rejected.

The adjusted p-value (stagewise-ordering p-value) is calculated from (9.7):

$$\begin{aligned} p_{adj} &= \alpha_1 + t(\beta_1 - \alpha_1) - \frac{1}{2}(\beta_1^2 - \alpha_1^2) \\ &= 0.1676(0.2) - \frac{1}{2}(0.2^2) = 0.0135. \end{aligned}$$

The confidence limit is calculated using (9.10). It is obvious that $\delta_{\alpha_1} = -\infty$. To obtain δ_{α_2}, we need to solve the following equation:

$$2 - \Phi\left(\frac{(0.055 - \delta_{\alpha_2})\sqrt{150}}{\sqrt{0.31}}\right) - \Phi\left(\frac{(0.045 - \delta_{\alpha_2})\sqrt{400}}{\sqrt{0.316}}\right) = 0.225 \quad (9.20)$$

Using the trial-error method in (9.20), we obtain $\delta_{\alpha_2} = 0.007$. Therefore the confidence limit is given by $\delta_{0\,\min} = 0.007$.

The adjusted estimate of treatment difference in event rate can be obtained from (9.18) with $w_i = \sqrt{0.5}$:

$$\delta_u = \frac{0.055\sqrt{0.5}\sqrt{150} + 0.045\sqrt{0.5}\sqrt{400}}{\sqrt{0.5}\sqrt{150} + \sqrt{0.5}\sqrt{400}} = 0.0488.$$

The unbiased estimate of treatment difference in event rate can be obtained from (9.15) with $w_1 = \frac{2(150)}{300+550} = \frac{6}{17}$ and $w_2 = 11/17$ from (9.16):

$$\delta_u = \frac{6(0.055)}{17} + \frac{11(0.045)}{17} = 0.0485.$$

The naive estimate is given by

$$\hat{\delta} = \frac{150(0.055)}{550} + \frac{400(0.045)}{550} = 0.0477$$

In summary, the test drug has about a 4.8% reduction in event rate with a one-sided 97.5% confidence limit of 0.7%. The adjusted p-value (one-sided) is 0.0135.

So far the binding futility boundaries are used. For the results with non-binding futility rule, see Problem 9.6.

Note that the actual trial was designed without SSR. The sponsor asked for SSR at interim analysis, and was rejected by the FDA. The trial eventually failed to demonstrate statistical significance.

Example 9.2: Adaptive Design with Farrington-Manning NI Margin

There are two ways to define the non-inferiority margin: (1) a prefixed non-inferiority (NI) margin and (2) a non-inferiority margin that proportional to the effect of the active control group, i.e., Farrington-Manning non-inferiority margin. The former has been discussed in Chapters 2 and 4. We now discuss the latter. The Farrington-Manning non-inferiority test was proposed for a classic design with a binary endpoint (Farrington and Manning, 1990), but can be extended to adaptive designs with different endpoints, in which the null hypothesis can be defined as $H_{ok} : u_{tk} - (1 - \delta_{NI})\, u_{ck} \leq 0$ for the k^{th} stage, where $0 < \delta_{NI} < 1$, u_{tk} and u_{ck} are the responses (mean, proportion, median survival time) for test and control groups at the k^{th} stage, respectively. The test statistic is defined as

$$T_k = \frac{u_{tk} - (1 - \delta_{NI})\, u_{ck}}{\sqrt{\sigma_t^2 + (1 - \delta_{NI})\sigma_c^2}}, \tag{9.21}$$

where $\sigma_t^2 = var\,(u_{tk})$ and $\sigma_c^2 = var\,(u_{ck})$ are given by Table 2.1 and (9.3) for a large sample. It is important to know that there is a variance reduction in comparison with the prefixed NI margin approach, in which the variance is $\sigma_t^2 + \sigma_c^2$ instead of $\sigma_t^2 + (1 - \delta_{NI})\sigma_c^2$. Therefore, Farrington-Manning test is usually much more powerful than the fixed margin approach. The SAS Macro 9.1 provides the capability for simulating both NI test methods with either balanced or imbalanced designs (See Problem 9.4).

9.7 Summary and Discussion

In this chapter, we have studied the different SSR approaches, in which we have combined the general adaptive design methods (MSP, MPP, and MINP/LW) with two sample-size adjustment rules given in (9.4) and (9.6). We have compared the performances of different approaches using the simulations. You can use the SAS Macro 9.1 or R program in the appendix to conduct your own simulations. In fact, it is strongly suggested to do so before selecting a design. Here is a summary of what we have studied in this chapter:

(1) Futility stopping can reduce the sample-size when the null is true. The futility boundary is suggested because in the case of a very small effect size, to continue the trial will require an unrealistically large sample-size, and an increased sample-size to N_{max} still may not have enough power.

(2) From an α-control point of view, for the methods in this chapter (MSP, MPP, and L-W), the algorithm for sample-size adjustment does not have to be predetermined; instead, it can be determined after we observed the results from the interim analysis.

(3) It might be a concern if the sample-size is based on a predetermined algorithm, IDMC's determination of the new sample-size will actually require the disclosure of the efficacy information for the trial. There are at least two approaches to handle this: (a) Use a discrete increment of sample-size; or (2) Set a target conditional power, and let the IDMC choose the new sample-size approximating to the conditional power with consideration of other factors.

(4) Unbiased estimates (point and confidence interval) can be obtained by using the fixed weight method, but interpretation may be difficult. In addition to the unbiased estimate, report the unbiased estimate with a bias assessment at the true mean = naive estimate.

(6) Power is an estimation, made at the initial design stage, of the probability of rejecting the null hypothesis. Therefore, it is less important when the sample-size can be adjusted at IA. In other words, the initial total sample-size is irrelevant to the final sample-size (it is only relevant for budgeting and operational planning), but the sample-size at the first stage is relevant.

(7) For adaptive designs, the conditional power is more important than the unconditional power. MSP often show superior over other methods. Non-binding futility rule is currently adopted by regulatory bodies. With non-binding futility boundaries, MSP is often superior over other methods.

It is interesting to know that increasing sample-size when the unblinded interim result is promising will not inflate the type-I error rate. The unblinded interim result is considered promising if the conditional power is greater than 50% or, equivalently, the sample-size increment needed to achieve a desired power does not exceed an upper bound (Chen, DeMets, and Lan, 2004).

There are other practical issues, e.g., what if the stagewise p_1 and p_2 are very different? Does this inconsistency cause any concern? If the answer is "yes," then should we check this consistency for a classic design too? MSP emphasizes the consistency, but MPP and MINP don't. More discussions on the controversial issues will be presented in Chapter 18.

Finally, it is interesting to know that a group sequential design with early stopping is a specific use of discrete functions of sample-size re-estimation.

Problems

9.1 Use the inverse-normal method to redesign the trial in Example 9.1.

9.2 Prove that the estimator given by Equation (9.19) is median unbiased, but that the mean bias and variance of $\hat{\delta}_u$ depend on the adaptation rule.

9.3 Proof: The maximum bias for a two-stage adaptive design with early stopping and SSR is bounded by

$$0.4\frac{\sigma}{\sqrt{n_1}}\frac{N_{\max}-n_1}{N_{\max}} = \frac{0.4\sigma}{\sqrt{n_1}}\frac{r_{\max}-1}{r_{\max}} < \frac{0.4\sigma}{\sqrt{n_1}},$$

where $N_{\max}$ is the maximum sample-size per group and $r_{\max} = N_{\max}/n_1$.

9.4 Study non-inferiority adaptive designs with both prefixed non-inferiority margin and Farrington-Manning margin and compare the results using both MSP and MLP with stopping boundaries in Table 7.5, where $w_1 = \sqrt{1/3}$ and $w_2 = \sqrt{2/3}$.

9.5 Design adaptive trial in Example 9.1 using MPP and MINP.

9.6 Design adaptive trial in Example 9.1 using non-binding futility rule.

9.7 Discuss the confidence interval defined by

$$\begin{cases} \delta_{0\min} = \max\left\{\delta_{\alpha_1}, ..., \delta_{\alpha_{K_s}}\right\} \\ \delta_{0\max} = \min\left\{\delta_{\beta_1}, ..., \delta_{\beta_{K_s}}\right\} \end{cases},$$

where

$$\begin{cases} \sum_{i=1}^{k}\left[1-\Phi\left(\frac{\hat{\delta}_i-\delta_{\alpha_k}}{\sigma}\sqrt{\frac{n_i}{2}}\right)\right] = \alpha_k \\ 1-\sum_{i=1}^{k}\left[1-\Phi\left(\frac{\hat{\delta}_i-\delta_{\beta_k}}{\sigma}\sqrt{\frac{n_i}{2}}\right)\right] = \beta_k \end{cases}, \ k = 1, ..., K_s,$$

Chapter 10

Multiple-Endpoint Adaptive Design

Multiple endpoints are common in clinical trials. In this chapter, we first discuss the multiple endpoint issues and commonly used methods in classic designs. We will review the single-stage and stepwise methods. The gatekeeper method is particularly interesting; we will extend this method to multiple-endpoint adaptive trials to develop the "fractal gatekeeper" method. A trial example will be used to illustrate step by step how to use the fractal gatekeeper method for analyzing multiple-endpoint adaptive trials.

10.1 Multiplicity Issues

It is well known that multiple analyses can inflate the type-I error dramatically without proper adjustment for the p-values or significance level. This is referred to as a multiplicity issue. Multiple analyses are sometimes necessary and the multiplicity can come from different sources: (1) multiple-treatment comparison, (2) multiple time point, (3) multiple endpoints, (4) multiple populations with same treatment, and (5) a combination of above sources.

Multiple-treatment comparisons are often conducted in dose-finding studies. Multiple time point analyses are often conducted in longitudinal studies with repeated measures or trials with group sequential or adaptive designs. We will focus on multiple-endpoint issues in this chapter; the multiple population issues will be discussed in Chapter 12.

Why multiple endpoint analysis? Lemuel Moyé pointed out (2003, p.76): There are three primary reason why we conduct a multiple endpoint study: (1) Diseases of unknown aetiologies or for which no clinical consensus on the single-most important clinical efficacy endpoint exist, (2) Diseases that manifest themselves in multi-dimensional ways, and (3) Therapeutic areas for which the prevailing methods for assessment of treatment efficacy dictate

a multi-faceted approach both for selection of the efficacy endpoints and their evaluations.

The statistical analyses of multiple endpoint problems can be categorized as (1) single primary efficacy endpoint with one or more secondary endpoints, (2) coprimary endpoints (more than one primary endpoint) with secondary endpoints, (3) composite primary efficacy endpoint with interest in each individual endpoint, (4) surrogate primary with supportive secondary endpoints. A surrogate endpoint is a biological or clinical marker that can replace a gold standard endpoint such as survival.

In the case of diseases of unknown etiologies where no clinical consensus has been reached on the single most important clinical efficacy endpoint, coprimary endpoints may be used. When diseases manifest themselves in multi-dimensional ways, drug effectiveness is often characterized by the use of composite endpoint, or global disease scores, or the disease activity index (DAI). When a composite primary efficacy endpoint is used, we are often interested in which particular aspect or endpoint the drug has demonstrated benefits. ICH Guideline suggests: "If a single primary variable cannot be selected from multiple measurements associated with the primary objective, another useful strategy is to integrate or combine the multiple measurements into a single or 'composite' variable, using a predefined algorithm.... This approach addresses the multiplicity problem without requiring adjustment to the type-I error." For some indications, such as oncology, it is difficult to use a gold standard endpoint, such as survival, as the primary endpoint because it requires longer follow-up time and because patients switch treatments after disease progression. Instead, a surrogate endpoint, such as time-to-progression, might be chosen as the primary with other supporting efficacy evidences, such as infection rate or time-to-skeleton event.

10.1.1 *Statistical Approaches to the Multiplicity*

Statistical approaches to multiplicity can be categorized as single-step and stepwise procedures and gatekeeper procedures – very special stepwise procedures. However, first let's discuss some basic concepts in the multiplicity discipline.

Basic Concepts

Error Inflation: Suppose we have two primary endpoints in a two-arm, active-controlled, randomized trial, and efficacy of the drug will be claimed as long as one of the endpoints is statistically significant. In such scenario, the familywise error (FWE), i.e., the probability of claiming efficacy when in fact there no efficacy will be inflated. The level of inflation is dependent on

the correlation between the two test statistics (Table 10.1). The maximum error rate inflation occurs when the endpoints are independent. If the two endpoints are perfectly correlated, there is no alpha inflation. For a correlation as high as 0.75, the inflation is still larger than 0.08 for a level 0.05 test. Hence, to control overall α, an alpha adjustment is required for each test. Similarly, to study how the alpha-inflation is related to the number of analyses, simulations are conducted; results are presented in Table 10.2. We can see that alpha is inflated from 0.05 to 0.226 with 5 analyses and to 0.401 with 10 analyses.

Table 10.1: Error Inflation Due to Correlations Between Endpoints

Level α_A	Level α_B	Correlation R_{AB}	Level α_{AB}
0.05	0.05	0	0.098
		0.25	0.097
		0.50	0.093
		0.75	0.083
		1.00	0.050

Note: $\alpha_A = \alpha$ for endpoint A, $\alpha_A = \alpha$ for endpoint B, and $\alpha_{AB} = \alpha$ for endpoint A or B.

Table 10.2: Error Inflation Due to Different Number of Endpoints

Level α_A	Level α_B	Number of analyses	Level α_{AB}
0.05	0.05	1	0.050
		2	0.098
		3	0.143
		5	0.226
		10	0.401

Familywise error control: There are two type of familywise type-I error controls: strong and weak. The strong control requires a α-level control for all possible null configurations (negation of the alternative hypothesis). A weak control requires a α-level control for the global null configuration but not necessarily all other null configurations.

*Closed family***:** A closed family is one for which any subset intersection hypothesis involving members of the family of tests is also a member of the family. For example, a closed family of three hypotheses H_1, H_2, H_3 has a total of 7 members, listed as follows: H_1, H_2, H_3, $H_1 \cap H_2$, $H_2 \cap H_3$, $H_1 \cap H_3$, $H_1 \cap H_2 \cap H_3$.

Closure Principle: Developed by Peritz (1970) and Marcus, Peritz, and Gabriel (1976). This principle asserts that one can ensure strong control of

FWE, and coherence (see below) at same time, by conducting the following procedure: Test every member of the closed family using a level α test (here, α refers to the comparison-wise error rate, not the FWE rate). A hypothesis can be rejected provided (1) its corresponding test was significant at α-level, and (2) every other hypothesis in the family that implies it has also been rejected by its corresponding α-level test.

Closed Testing Procedure: A test procedure is said to be closed if and only if the rejection of a particular univariate null hypothesis at a given significance α-level implies the rejection of all higher level (multivariate) null hypotheses containing the univariate null hypothesis at the same α-level. The procedure can be described as follows (Bretz, et al., 2006):

(1) Define a set of elementary hypotheses $H_1; ...; H_n$ of interest.

(2) Construct all possible $m > n$ intersection hypotheses $H_I = \cap H_i$, $I \subseteq \{1, ..., n\}$.

(3) For each of the m hypotheses find a suitable local level-α test.

(4) Reject H_i at FWE rate α, if all hypotheses H_I with $i \in I$ are rejected, each at (local) level α.

This procedure specifically is not often used in practice. However, the closure principle has been used to derive many useful test procedures such as those of Holm (1979), Hochberg (1988), Hommel (1988), the multivariate Hotelling's T_2 test and the family of direction-sensitive linear combination test statistics of O'Brien types can be closed by direct application of the closure principle.

Partition Principle: Similar to the closed testing procedure, strong control over the familywise α-level for the null hypotheses is formed by partitioning the parameter space into disjointed partitions with some logical ordering. Tests of the hypotheses are carried out sequentially at different partitioning steps. The process of testing stops upon failure to reject a given null hypothesis for predetermined partitioning steps.

Coherence and *consonance* are two interesting concepts. *Coherence:* If hypothesis H implies H^*, then whenever H is retained, so must be H^*. *Consonance:* Whenever H is rejected, at least one of its components is rejected too. Coherence is a necessary property of closed testing procedures; consonance is desirable but not necessary. A procedure can be coherent but not consonant because of asymmetry in the hypothesis testing paradigm. When H is rejected, we conclude that it is false. However, when H is retained, we do not conclude that it is true; rather, we say that there is no sufficient evidence to reject it. Multiple comparison procedures that satisfy the closure principle are always coherent but not necessarily consonant (Westfall, et al., 1999).

10.1.2 *Single Step Procedures*

The commonly used single-stage procedures include the Sidak method with strong FWE control under independence (Sidak 1967), the simple Bonferroni method with strong FWE control, the Simes-Bonferroni method (Global test, Simes 1986), and Dunnett's test for all active arms against the control (Dunnett 1955). In the single-step procedure, to control the FWE, the unadjusted p-values are compared against the adjusted alpha to make the decision to reject or not reject the corresponding null hypothesis. Alternatively, we can use adjusted p-value to compare against the original α for the decision-making.

Sidak method

The Sidak method is derived based on the simple fact that the probability rejecting at least one null hypothesis is equal to 1-Pr(all null hypotheses are correct). To control FWE, the adjusted alpha α_k for the null hypothesis H_{ok} can be found by solving the following equation:

$$\alpha = 1 - (1 - \alpha_k)^n . \tag{10.1}$$

Therefore, the adjusted alpha is given by

$$\alpha_k = 1 - (1 - \alpha)^{1/n}, \ k = 1, 2, ..., n. \tag{10.2}$$

If p-value is less than or equal to α_k, reject H_{ok}. Alternatively we can calculate the adjusted p-value:

$$\tilde{p}_k = 1 - (1 - p_k)^n . \tag{10.3}$$

If the adjusted p-value $\tilde{p}_k$ is less than α, then reject H_{ok}.

Simple Bonferroni method

Simple Bonferroni method is a simplification of Sidak method by using Bonferroni inequality:

$$P(\cup_{k=1}^{n} H_k) \leq \sum_{k=1}^{n} P(H_k). \tag{10.4}$$

Based on (10.4), we can conservatively use the adjusted alpha:

$$\alpha_k = \frac{\alpha}{n} \tag{10.5}$$

and the adjusted p-value:

$$\tilde{p}_k = np_k. \tag{10.6}$$

This is a very conservative approach without consideration of any correlations.

Simes-Bonferroni method

This is a test for the global null hypothesis, i.e., the FWE is weakly controlled. Using the Simes-Bonferroni method, we reject the null hypothesis if

$$p_k \leq \frac{k\alpha}{n}. \tag{10.7}$$

Therefore, the adjusted alpha is given by

$$\alpha_k = \frac{k\alpha}{n},$$

and the adjusted p-value is given by

$$\tilde{p}_k = \frac{n}{k}p_k.$$

Dunnett's method

Dunnett's method is used for multiple comparisons of active groups against a common control group, which is often seen in clinical trials with multiple parallel groups. Let n_0 and n_i $(i = 1, ..., m)$ be the sample-size for the control and the i^{th} group, the test statistic is given by (Westfall, et al., 1999, p.77)

$$T^m = \max_i \frac{\bar{y}_i - \bar{y}_0}{\sigma\sqrt{1/n_i + 1/n_0}}, \tag{10.8}$$

$$P\left(T^m \leq c\right) = \int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^m \left\{ \Phi\left[\frac{\lambda_i z + cu}{\left(1-\lambda_i^2\right)^{1/2}}\right] - \Phi\left[\frac{\lambda_i z - cu}{\left(1-\lambda_i^2\right)^{1/2}}\right] \right\} d\Phi\left(z\right) dF_{df}\left(u\right), \tag{10.9}$$

where

$$\frac{d\Phi\left(z\right)}{dz} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \tag{10.10}$$

is the standard normal density function and

$$\frac{dF_v(u)}{du} = \frac{v^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right) 2^{\frac{v}{2}-1}} u^{v-1} e^{-\frac{vu^2}{2}} \tag{10.11}$$

is the density of $\sqrt{V/v}$, where V is Chi-squared random variable with v degrees of freedom. The parameter λ_i is given by

$$\lambda_i = \left(\frac{n_i}{n_0 + n_i}\right)^{1/2}. \tag{10.12}$$

The calculation of (3.9) requires numerical integrations (Hochberg and Tamhane, 1987, p.141).

Fisher-Combination Test

To test the global null hypothesis $H_o = \cap_{i=1}^{m} H_{oi}$, we can use the so-called Fisher combination statistic:

$$\chi^2 = -2\sum_{i=1}^{m} \ln(p_i), \tag{10.13}$$

where p_i is the p-value for testing H_{oi}. When H_{oi} is true, p_i is uniformly distributed over [0,1]. Further more if p_i $(i = 1, ..., m)$ are independent, the test statistic χ^2 is distributed as a chi-square statistic with $2m$ degrees of freedom, thus H_o is rejected if $\chi^2 \geq \chi^2_{2m,1-\alpha}$. Note that if p_i is not independent or H_o is not true (one of H_{oi} is not true), then χ^2 is not necessarily a chi-square distribution. Therefore, Fisher combination method is weakly controlled for α.

10.1.3 *Stepwise Procedures*

Stepwise procedures are different from single-step procedures, in the sense that a stepwise procedure must follow a specific order to test each hypothesis. In general, stepwise procedures are more powerful than single-step procedures. There are three categories of stepwise procedures that are dependent on how stepwise tests proceed: stepup, stepdown, and fixed sequence procedures. The commonly used the stepwise procedures include Bonferroni-Holm stepdown method (Holm 1979), Sidak-Holm stepdown method (Westfall, et al., 1999, p.31), Hommel's procedure (Hommel 1988), Hochberg's stepup method (1990), Rom's method (1990), and sequential test with fixed sequences (Westfall, et al., 1999).

Step-down Procedure (From the most significance to the least significance):

In a stepdown procedure, the p-value is arranged in an ascending order, i.e., from the smallest to the largest:

$$P_{(1)} \leq P_{(2)} \leq ... \leq P_{(n)} \tag{10.14}$$

with the corresponding hypotheses

$$H_{(1)}, H_{(2)}, ..., H_{(n)}.$$

The test is proceed from $H_{(1)}$ to $H_{(n)}$. If, $P_{(k)} > C_k\alpha$ $(k = 1, ..., n)$, retain all $H_{(i)}$ $(i \geq k)$; otherwise, reject $H_{(k)}$ and continue to test $H_{(k+1)}$. The critical values C_k are different for different procedures.

The adjusted p-values are

$$\begin{cases} \tilde{P}_{(1)} = C_1 P_{(1)} \\ \tilde{P}_{(k)} = \max\left(\tilde{P}_{(k-1)}, C_k P_{(k)}\right), \ k > 1, ..., n. \end{cases} \tag{10.15}$$

Therefore an alternative test procedure is to compare the adjusted p-values against the unadjusted α. After adjusting p-values, one can test the hypotheses in any order.

For the Bonferroni-Holm stepdown procedure, $C_k = n - k + 1$. Hommel procedure (Hommel, 1988; Westfall, et al., 1999) derived from Closure Principle is specified as follows: compute $j = \max k \in \{1, ..., n\}$ subject to $P_{n-k+i} > \frac{i\alpha}{k}$, $i = 1, .., k$. If the maximum does not exist, reject all H_k, $k = 1, ..., n$; otherwise reject H_i with $p_k \leq \alpha/j$.

Stepup Procedure (From the least significance to the most significance)

The stepup procedure proceeds from $H_{(n)}$ to $H_{(1)}$. If, $P_{(k)} \leq C_k\alpha$ $(k = 1, ..., n)$, reject all $H_{(i)}$ $(i \leq k)$, otherwise, retain $H_{(k)}$ and continue to test $H_{(k-1)}$. The critical values C_k for Hochberg stepup procedure is $C_k = n - k + 1$ $(k = 1, .., n)$.

The adjusted p-values are

$$\begin{cases} \tilde{P}_{(n)} = C_n P_{(n)}, \\ \tilde{P}_k = \min\left(\tilde{P}_{(n-k+1)}, C_k P_{(k)}\right), \ k > 1, ..n. \end{cases} \tag{10.16}$$

Therefore, an alternative test procedure is to compare the adjusted p-values against the unadjusted α.

Hochberg stepup method does not control FWE for all correlations, but it is a little conservative when p-values are independent (Westfall, et al. 1999, p.33). Rom (1990) method controls α exactly for independent p-values. However, the calculation of C_k is complicated.

Sequential test with fixed sequences

This procedure is a stepdown procedure with the order of hypotheses predetermined:

$$H_1, H_2, ..., H_n.$$

The test is proceed from H_1 to H_n. If, $P_k > \alpha$ $(k = 1, ..., n)$, retain all H_i $(i \geq k)$. Otherwise, reject H_k and continue to test H_{k+1}.

The adjusted p-values are

$$\begin{cases} \tilde{P}_1 = P_1 \\ \tilde{P}_k = \max\left(\tilde{P}_{k-1}, P_k\right), \ k > 1, .., n. \end{cases} \tag{10.17}$$

The sequence of the tests can be based on the importance of hypotheses or the power of the tests. Note that if the previous test is not significant, the next test will not proceeds even if its p-value is extremely small.

Benjamini-Hochberg procedure for FPR

The Benjamini-Hochberg procedure is taken the stepup fashion with $C_k = \frac{k-1}{k}\alpha$. The adjusted p-values are given by (10.15) with the new C_k, where FPR (False positive rate) is defined as the expected proportion of erroneously rejected null hypotheses among the rejected ones (Benjamini and Hochberg, 1995; Westfall, et al., 1999, p.21).

Other methods such as Tukey and Scheffe methods can be found in many multiplicity literatures, we are not going to discuss here.

10.1.4 *Gatekeeper Approach*

Gatekeeper approach (Dmitrienko, et al., 2005, p.106-127) is an extension of the fixed sequence method. The method is motivated by the following hypothesis testing problems in clinical trials. (1) Benefit of secondary endpoints can be claimed in the drug label only if the primary endpoint is statistically significant. (2) If there are coprimary endpoints (multiple primary endpoints), secondary endpoints can be claimed only if one of the

primary endpoints are statistically significant. (3) In multiple-endpoint problems, the endpoints can be grouped based on their clinical importance.

Suppose there are n null hypotheses to test. We group them into m families. Each family is a composition of hypotheses. The hypotheses in the $i^{th}(i = 1, ..., m_i)$ family are denoted by either serial gatekeeper

$$F_i = H_{oi1} \cup H_{oi2} \cup ... \cup H_{oim_i} \tag{10.18}$$

or parallel gatekeeper

$$F_i = H_{oi1} \cap H_{oi2} \cap ... \cap H_{oim_i} \tag{10.19}$$

or mixed type

$$F_i = (H_{oi1} \cup ... \cup H_{oim_i}) \cup (H_{oi1} \cap ... \cap H_{oin_i}) . \tag{10.20}$$

The hypothesis testing proceeds from the first family F_1 to the last family F_m. To test F_i $(i = 2, ..., m)$, the test procedure has to pass $i - 1$ previous gatekeepers, i.e., reject all F_k $(k = 1, ..., i - 1)$ at predetermined level of significance α.

For parallel gatekeeper we can either weakly or strongly control family-wise type-I error. For serial gatekeeper, we always strongly control the family-wise error.

In the current regulatory settings, the primary endpoint provides the basis for approval of an investigational drug, while a significant improvement in a secondary endpoint is not generally considered as substantial evidence of therapeutic benefit (O'Neill, 1997). The secondary endpoints can be (CPMP, 2002): (1) variables that may potentially provide the basis for a new regulatory claim, (2) variables that may become the basis for additional claims, (3) variables yielding supportive evidence.

Example 10.1 Acute Respiratory Disease Syndrome Trial

Suppose a parallel, placebo-controlled trial in patients with acute respiratory disease syndrome (ARDS). In this 28-day study, the coprimary endpoints are ventilation-free days (VFD) and 28-day mortality (Death); the secondary endpoints are the number of ICU-free days (ICUFD, out of the intensive care unit) and general quality of life (QOL). The secondary endpoints are used for additional regulatory claim in the label.

Suppose the gatekeeper procedure is used. Specifically, VFD and Death are tested separately at one-sided $\alpha = 0.0125$. If either is significant, the

null will be rejected and drug efficacy is claimed. Furthermore, the secondary endpoints, ICUFD and QOL will be tested separately at one-sided $\alpha = 0.0125$. If either or both endpoints are significant, the benefit on the corresponding secondary endpoints will be claimed. On the other hand, if one-sided p-values for both VFD and Death are larger than 0.0125, then no efficacy will be claimed and no further test on the secondary endpoints will be pursued.

10.2 Multiple-Endpoint Adaptive Design

As we discussed earlier, a clinical trial usually involves multiple endpoints. If there are more than two primary endpoints, they are called coprimary endpoints. However, more often a trial has one primary and several secondary endpoints. In the coprimary case, as long as the treatment effect is statistically significant on one of the primary endpoints, the efficacy criterion is met statistically. On the other hand, for the case of primary-secondary endpoint, the statistical significance must be achieved on the primary endpoint before the significance of any of the secondary endpoints can be claimed on the drug label. In this section, we will discuss sequential adaptive designs for multiple-endpoint trials. Kieser (1999) proposed a test procedure, in which he considered multiple-endpoint test at each time-point and multiple-endpoint adjustments are made in the same way as classic design based on α spent on each time point (e.g., α_1 for the first interim analysis). In contrast, Tang and Geller (1999) proposed a different approach for classic group sequential design, in which, they view different endpoints hierarchically and fixed sequence tests are constructed based on the importance of the endpoints. The Tang-Geller's method is more powerful than Kieser's method. Chang (2007) proposed a procedure to extend the Tang-Geller's method. The method can be used for adaptive designs and is even more powerful than Tang-Geller's method.

10.2.1 *Fractals of Gatekeepers*

Recall that the gatekeeper procedure can be described as follows: The hypothesis testing proceeds from the first family F_1 to the last family F_m, to test F_i $(i = 2, ..., m)$, the test procedure has to pass $i-1$ previous gatekeepers, i.e., reject all F_k $(k = 1, ..., i-1)$ at predetermined level of significance α. We can extend the family structure to multiple hierarchical levels.

The fractals of gatekeepers refer to the nested gatekeepers (or family of

hypotheses) at different levels. For example:

$$F_i = \cup_{j=1}^{m_i} F_{ij} \tag{10.21}$$

and F_{ij} itself is a family of hypotheses, i.e., serial gatekeepers

$$F_{ij} = \cap_{k=1}^{n_{ij}} F_{ijk} \tag{10.22}$$

or parallel gatekeepers

$$F_{ij} = \cup_{k=1}^{n_{ij}} F_{ijk} \tag{10.23}$$

The hypothesis (family) F_{ijk} can further be parallel or serial gatekeepers, and this process can continue even further. For convenience, we call F_i, F_{ij}, and F_{ijk} the first level, second, and third level families, respectively, dependent on their number of subscripts.

In general, the hypotheses from the same family or different families are not independent The correlation between two hypotheses (or families) are usually very complicated. Therefore, conservative approaches can be used, in which no correlations are considered. For the hypothesis families at each level, we either use the single-step Bonferroni approach or the fixed sequence procedure such that the FWE is strongly controlled.

A general test procedure is described as follows: At the first level of the hypothesis family with m_1 subfamilies, the familywise α-level is α. If their subfamilies are parallel-structured, we use parallel gatekeeper approach, i.e., test each subfamily at the level of $\alpha_{F_1} = \alpha/m_1$. If these subfamilies are serial-structured, we use serial gatekeeper approach, i.e., test each subfamily at the level of $\alpha_{F_1} = \alpha$ with a fixed sequence. At the second family level with m_2 subfamilies, the familywise α-level is α_{F_1}. Again if their subfamilies are parallel-structured, we use the parallel gatekeeper approach, i.e., test each subfamily at the level of $\alpha_{F_2} = \alpha_{F_1}/m_2$. If these subfamilies are serial-structured, we use serial gatekeeper approach, i.e., test each subfamily at the level of $\alpha_{F2} = \alpha$ with a fixed sequence. This process continues until the α-level for each individual hypothesis test (not hypothesis test family) is determined. The actual test procedure proceeds from the lowest level (i.e. individual hypothesis level) to the highest level of hypothesis family, at which the α-level is α.

We illustrate this method with two different scenarios in an adaptive design: (1) one primary and several secondary endpoints, (2) coprimary and several endpoints.

10.2.2 *Single Primary with Secondary Endpoints*

Example 10.2 Three-Stage Adaptive Design for NHL Trial

A phase-III two parallel group Non-Hodgkin's Lymphoma trial was designed with three analyses. The primary endpoint is progression-free survival (PFS), the secondary endpoints are (1) overall response rate (ORR) including complete and partial response and (2) complete response rate (CRR). The estimated median PFS is 7.8 months and 10 months for the control and test groups, respectively. Assume a uniform enrollment with an accrual period of 9 months and a total study duration of 23 months. The estimated ORR is 16% for the control group and 45% for the test group. The classic design with a fixed sample-size of 375 subjects per group will allow for detecting a 3-month difference in median PFS with 82% power at a one-sided significance level of $\alpha = 0.025$. The first interim analysis (IA) will be conducted on the first 125 patients/group (or total $N_1 = 250$) based on ORR. The objective of the first IA is to modify the randomization. Specifically, if the difference in ORR (test-control), $\Delta_{ORR} > 0$, the enrollment will continue. If $\Delta_{ORR} \leq 0$, then the enrollment will stop. If the enrollment is terminated prematurely, there will be one final analysis for efficacy based on PFS and possible efficacy claimed on the secondary endpoints. If the enrollment continues, there will be an interim analysis based on PFS and the final analysis of PFS. When the primary endpoint (PFS) is significant, the analyses for the secondary endpoints will be performed for the potential claim on the secondary endpoints. During the interim analyses, the patient enrollment will not stop. The number of patients at each stage is approximately as shown in Figure 10.1.

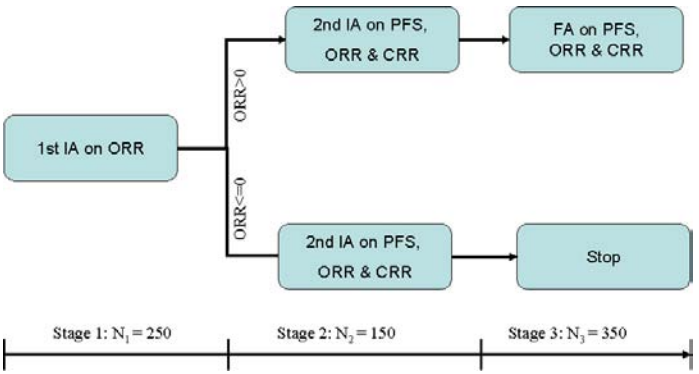


Figure 10.1: Multiple-Endpoint Adaptive Design

We use two different adaptive methods for this designs: MINP and MSP as illustrated below.

I. Inverse-normal Method

The test statistic at the k^{th} analysis is defined as:

$$z_{ik} = \sum_{j=1}^{k} w_{kj} \Phi^{-1}(1 - p_{ij}), \tag{10.24}$$

where the subscript i represents the i^{th} endpoint, i.e., 1 for PFS, 2 for ORR, 3 for CRR. p_{ij} is the stagewise p-value for the i^{th} endpoint based on the subsample from the j^{th} stage.

$$w_{kj} = \sqrt{\frac{N_k}{\sum_{j=1}^{k} N_j}}. \tag{10.25}$$

The first IA is not intent to claim efficacy or futility, but modify the enrollment (continue or not continue enrollment) based ORR. The stagewise test statistic is given by

$$z_{i1} = (1 - p_{i1}). \tag{10.26}$$

At the 2^{nd} IA, the test statistic is given by

$$Z_{i2} = 0.8\Phi^{-1}(1 - p_{i1}) + 0.6\Phi^{-1}(1 - p_{i2}). \tag{10.27}$$

If the trial continues after the 2^{nd} IA, the test statistic at final analysis will be

$$Z_{i3} = 0.58\Phi^{-1}(1 - p_{i1}) + 0.45\Phi^{-1}(1 - p_{i2}) + 0.68\Phi^{-1}(1 - p_{i3}). \tag{10.28}$$

The OB-F stopping boundaries on the z-scale are $\alpha_1 = 3.490$, $\alpha_2 = 2.468$, $\alpha_3 = 2.015$ for stage 1, 2, and 3, respectively (Table 10.3). For simplicity, the same stopping boundaries are used for PFS, ORR, and CR.

Denote H_{oij} the null hypothesis for the i^{th} endpoint at j^{th} stage. The gatekeeper test procedure is described as follows: Construct the first hypothesis family as $F_1 = H_{o11} \cap H_{o12} \cap H_{o13}$ for the PFS, similarly $F_2 = H_{o21} \cap H_{o22} \cap H_{o23}$ for the ORR, and $F_3 = H_{o31} \cap H_{o32} \cap H_{o33}$. F_1 is tested at level α, if F_1 is not rejected, no further test will be tested. If F_1 is rejected, we further test F_2. If F_2 is not rejected, no further test will proceed. If F_2 is rejected, F_3 is tested. All tests will be conducted at the same level $\alpha = 0.025$. The (closed set) gatekeeper procedure ensures the strong control of FWE. Note that due to the correlation between PFS

and ORR, we can not just consider a two-stage weighting test statistic for PFS, even if the hypothesis test for PFS is not performed at the first IA.

Suppose at the first IA, the stagewise p-value for the primary endpoint (PFS) is $p_{11} = 0.030$ and the test statistic is $z_{11} = \Phi^{-1}(1 - p_{11}) = 1.881 < \alpha_1$, therefore the trial continues. At the second IA, the stagewise p-value $p_{12} = 0.034$, and the test statistic $z_{12} = 2.6$ is calculated from (10.27). Therefore the null hypothesis for PFS is rejected and trial stops. Because the PFS is rejected, we can now test ORR. Suppose the stagewise p-value for ORR is $p_{21} = 0.002$ and $z_{21} = \Phi^{-1}(1 - p_{21}) = 3.60 > 3.490$; hence the null hypothesis for ORR is rejected. Because ORR is rejected, we can proceed with the test for CRR. Suppose that $p_{31} = 0.1$ and $z_{31} = 1.281 < \alpha_1$; $p_{32} = 0.12$ and $z_{32} = 1.73 (< \alpha_2)$ from (10.27). Due to rejection of PFS at the second IA, the trial was stopped. However, the enrollment was not stopped during the interim analyses. At the time when the decision was made to stop the trial, 640 patients (approximately 320 per group) were enrolled. The gatekeeper procedure allows us to proceed to the third analysis of CRR. However, the rejection boundary needs to be recalculated through numerical integration or simulation based on the OB-F spending function. Based on the information time $\sqrt{640/750} = 0.923\,76$, the new rejection boundary is approximately $\alpha^*_{33} = 2.10$. Suppose that the observed $p_{33} = 0.065$ and the test statistic $z_{33} = 2.3$ is calculated from (10.28). Therefore CRR is also rejected. We can see that PFS, ORR, and CRR were all rejected, but at different times!!! The closed test project allows the rejections of different endpoints at different times (IAs) as illustrated in this example. The calculation is summarized in Table 10.3.

Table 10.3: MINP Based on Hypothetical p_{ik}

		PFS		ORR		CRR	
IA (k)	α_k	p_{1k}	z_{1k}	p_{2k}	z_{2k}	p_{3k}	z_{3k}
1	3.490	0.030	1.881	0.0002	3.60	0.10	1.281
2	2.468	0.034	2.600			0.12	1.730
3	2.015					0.065	2.300

II. Recursive Two-stage Adaptive Design with MSP

We now use the recursive two-stage adaptive design (Chapter 8) as described in the following steps.

(1) Calculate the sample-size based on the classic fixed sample-size design.

For median times of 7.8 months and 10 months for the two groups, an enrollment duration of 9 months, and a trial duration of 23 months, a sample-size of 350/group will allow for detecting a 2.2-month difference in

PFS with approximately 80% power at one-sided $\alpha = 0.025$.

(2) Design the first two-stage trial.

The IA will be conducted based on a sample-size of $n_1 = 150$ subjects per group. Using MSP, we decide the stopping boundaries at IA: $\alpha_1 = 0.01$ and $\beta_1 = \alpha_2 = 0.1832$ from (8.49) or (4.8), we obtain:

$$\begin{aligned}\alpha_2 &= \sqrt{2(\alpha - \alpha_1)} + \alpha_1 \\ &= (2(0.025 - 0.01))^{0.5} + 0.01 \\ &= 0.1832\end{aligned}$$

(3) Conduct the first interim analysis.

Assume $p_{11} = 0.030 > \alpha_1$ and the null hypothesis for PFS is not rejected. The trial should proceed to the next stage.

(4) Decide whether to redesign a new two-stage.

To determine if we need to redesign a new two-stage trial, the conditional error is calculated (8.48) as $\pi_2 = \alpha_2 - p_{11} = 0.1832 - 0.03 = 0.153\,2$. To have 90% conditional power, a total of 241 subjects/group are required based on (8.51) or SAS Macro 7.2 (assumed the standard effect size = 0.21). We decided to add a new two-stage design, $n_{21} = 150$ for second IA (the first stage of the second two-stage design), which provides a conditional reject probability of 79%. The stopping boundaries for the second two-stage design are specified as follows: $\alpha_{21} = 0.15$, Let $\beta_{21} = \alpha_{22}$. From (8.49) or (4.8), we obtain:

$$\begin{aligned}\alpha_{22} &= \sqrt{2(\pi_2 - \alpha_{21})} + \alpha_{21} \\ &= (2(0.1532 - 0.15))^{0.5} + 0.15 \\ &= 0.23\end{aligned}$$

Note that we also need to determine the stopping boundaries for ORR and CRR, but they don't have to be the same, i.e., the stopping boundaries can be different for different endpoints. However, the new stopping boundaries have to be determined for all endpoints at each interim analysis regardless of the endpoint for which the IA was performed. For simplicity, we use the same stopping boundaries for all three endpoints.

(5) Perform the 2^{nd} interim analysis.

Assume $p_{12} = 0.04$. Therefore, the p-value for the PFS at the 2^{nd} IA is $p_{11} + p_{12} = 0.03 + 0.04 = 0.07 < \alpha_{21} = 0.15$. Hence the null hypothesis for the PFS is rejected.

(6) Because the null hypothesis family F_1 for PFS is rejected, we proceed

to the test for ORR. Assume $p_{21} = 0.007 < \alpha_1 = 0.01$. Hence the null hypothesis ORR is rejected.

(7) Because ORR is rejected, we proceed to the test for CRR. Assume $p_{31} = 0.05 > \alpha_1 = 0.01$ at the 1^{st} IA. At the 2^{nd} IA, the test statistic $T_2 = p_{31} + p_{32} = 0.05 + 0.11 = 0.16 > \alpha_{21} = 0.15$. At the last analysis, the test statistic for CRR is $T_2 = p_{31} + p_{32} + p_{33} = 0.05 + 0.11 + 0.05 = 0.21 < \alpha_{22} = 0.23$, therefore the null hypothesis for CRR is rejected.

Table 10.4: RTAD Based on Hypothetical p_{ik}

		PFS	ORR	CRR
IA (k)	α_k	p_{1k}	p_{2k}	p_{3k}
1	0.01	0.03	0.007	0.05
2	0.15	0.04		0.11
3	0.23			0.05

10.2.3 *Coprimary with Secondary Endpoints*

When there is more than one primary endpoint, the test of the primary can be either fixed sequence or parallel. If there are more than two primary endpoints or more than two secondary endpoints, the fractal gatekeeper method can be used.

Example 10.3 Design with Multiple Primary-Secondary Endpoints

Suppose we are interested in two primary and three secondary endpoints. There are two family members at the first level, corresponding to primary and secondary endpoints. We will call them primary and secondary families. With the primary family there are two individual hypotheses for the two primary endpoints. With the secondary family there are three individual hypotheses for the three endpoints.

Strategy 1:

We will test the two families with the serial-gatekeeper procedure, therefore the familywise α-level for both the primary and secondary families is α. Within the primary family, the two hypotheses will be tested sequentially with a fixed sequence; each test is performed at level α. Similarly, within the secondary family, the three hypotheses will be tested sequentially with a fixed sequence; each test is performed at level α.

Strategy 2:

We will test the two families with the serial-gatekeeper procedure, therefore the familywise α-level for both the primary and secondary families is α. Within the primary family, the two hypotheses will be tested sequentially with a fixed sequence; each test is performed at level α. However, within the

secondary, each of the three hypotheses will be tested at level $\alpha/3$ without specification of any order.

Strategy 3:

We will test the two families with the serial-gatekeeper procedure, therefore the familywise α-level for both the primary and secondary families is α. Within the primary family, each of the two hypotheses will be tested at level $\alpha/2$. However, within the secondary family, the three hypotheses will be tested sequentially with a fixed sequence at the same level α.

Strategy 4:

We will test the two families with the serial-gatekeeper procedure, therefore the familywise α-level for both the primary and secondary families is α. Within the primary family, each of the two hypotheses will be tested at level $\alpha/2$. Similarly, within the secondary family, each of the three hypothesis will be tested at level $\alpha/3$ without specification of any order.

Strategy 5:

We will test the two families with the serial-gatekeeper procedure, therefore the familywise α-level for both the primary and secondary families is α. Within the primary family, each of the two hypotheses will be tested at level $\alpha/2$. Within the secondary family, we further group the three hypotheses into two families: The first one has a single hypothesis test for the first secondary endpoint, and the second family consists of the hypotheses for the other two secondary endpoints. We test the two families sequentially with a fixed sequence, both at level α. However within the second family for the secondary endpoints, the two individual hypotheses will be tested in parallel. Therefore, the first secondary endpoint is tested at level α but the 2^{nd} and 3^{rd} secondary endpoints are tested at level $\alpha/2$. Note that due to the fractal gatekeeping method, if one of the hypotheses in the precedent family is not rejected, no hypothesis test will performed further.

There are even more possible strategies. See the general test procedure for the fractal gatekeeping approach.

10.2.4 *Tang-Geller Method*

Tang and Geller (1999) proposed the following test procedures to group sequential design with multiple endpoints. This method can be generalized to adaptive design, though their procedures are less powerful than the method in the previous section.

Let $M = \{1, 2, ..., m\}$ be the set of indices for the m endpoints. Let F denote a non-empty subset of M and $H_{o,F}$ the null hypothesis $u_i = 0$ for $i \in F$. Let T_F be a test statistic for $H_{o,F}$. Consider a group sequential trial with K analyses. We use $T_{F,t}$ to indicate the dependence of T_F on the

analysis time t. Let $\{\alpha_{F,t}, t = 1, 2, ..., K\}$ be a one-sided stopping boundary for testing $H_{o,F}$ such that $P_{Ho,F}\{T_{F,t} > \alpha_{F,t}$ for some $t\ \} \leq \alpha$. For a given vector u, let $I_u = \{i, u_i = 0\}$.

Tang and Geller proposed the following two procedures that preserve strong control of type-I error.

Procedure 1:

Step 1. Conduct interim analyses to test $H_{o,M}$, based on the group sequential boundary $\{\alpha_{M,t},\ t = 1, 2, ..., K\}$.

Step 2. When $H_{o,M}$ is rejected, say at time t^*, stop the trial and apply the closed testing procedure to test all the other hypotheses $H_{o,F}$ using T_{F,t^*} with α_{F,t^*} as the critical value.

Step 3. If the trial continues to the last analysis without rejection of $H_{o,M}$, then no hypotheses are rejected.

Proof: A type-I error occurs if, for some $F \subseteq I_u$, $H_{o,F}$ is rejected, where u denotes the underlying parameter vector (e.g., difference in mean, response rate, or median survival time). According to the closed testing procedure, $H_{o,F}$ can be rejected only if H_{o,I_u} is rejected. Thus, {type-I error occurs} $= \cup_{t=1}^{K}$\{type-I error occurs at time $t\ \} \subseteq \cup_{t=1}^{K}$ \{reject H_{o,I_u} at time $t\}$ $\subseteq \cup_{t=1}^{K}\{T_{I_u,t} > \alpha_{I_u,t}\} = \{T_{I_u,t} > \alpha_{I_u,t}$, for some t, $1 \leq t \leq K\}$. Hence, the probability of making at least one type-I error is at most $P\{T_{I_u,t} > \alpha_{I_u,t}$, for some $t\} \leq \alpha$.

Procedure 1 does not allow continuation of the trial once the global test crosses its boundary. Tang and Geller further developed Procedure 2 below, which allows the trial to continue until all hypotheses are rejected or the last analysis is conducted.

Procedure 2:

Step 1. Conduct interim analyses to test $H_{o,K}$, based on the stopping boundary $\{\alpha_{K,t}, t = 1, 2, ..., K\}$.

Step 2. When $H_{o,M}$ is rejected, say at time t^*, apply the closed testing procedure to test all the other hypotheses $H_{o,F}$ using T_{F,t^*} with α_{F,t^*} as the critical value.

Step 3. If any hypothesis is not rejected, continue the trial to the next stage, in which the closed testing procedure is repeated, with the previously rejected hypotheses automatically rejected without retesting.

Step 4. Reiterate step 3 until all hypotheses are rejected or the last stage is reached.

You may have noticed that the procedure we have used in Examples 10.2 and 10.3 can be described as follows.

Procedure 3 (Chang, 2007)**:**

Step 1. Conduct interim analyses to test $H_{o,K}$, based on the group

sequential boundary $\{\alpha_{K,t}, t = 1, 2, ..., K\}$.

Step 2. When $H_{o,M}$ is rejected, say at time t^*, apply the closed testing procedure to test all the other hypotheses $H_{o,F}$ using $T_{F,t^{**}}$ with $\alpha_{F,t^{**}}$ as the critical value for any predetermined IA time $t^{**} \leq t^*$.

Step 3. If any hypothesis is not rejected, continue the trial to the next stage, in which the closed testing procedure is repeated, with the previously rejected hypotheses automatically rejected without retesting.

Step 4. Reiterate step 3 until all hypotheses are rejected or the last stage is reached.

10.2.5 *Summary and Discussion*

In this chapter, we have introduced several important concepts, principles, and test procedures for multiplicity issues. The closure principle is very useful, from which many test procedures are developed including the stepup, stepdown, fixed sequence, and gatekeeper procedures. The fractal gatekeeper procedure is very general and can be applied to very complicated test scenarios. Following Tang and Geller's idea, we construct an improved closed test procedure for adaptive designs. There are other methods that construct test-family based on the analysis time-point. Those methods are less powerful.

Problem

10.1 Prove the fractal gatekeeper method preserve strong $\alpha-$control.

10.2 This is a case presented by FDA statisticians (Hung, O'Neill, Wang, and Lawrence, 2006). In an oncology clinical trial, time-to-disease progression (TTP) was a primary surrogate endpoint and overall survival (OS) was the ultimate primary efficacy endpoint. A group sequential design was employed with an alpha-spending function pre-specified for the TTP endpoint. After TTP was shown to be highly statistically significant at an interim analysis, the trial was stopped. An important question is how to test the other endpoint, OS, given that perhaps only 20% of the planned total number of deaths is available. One suggestion is to use the same alpha-spending function as used for TTP or an alpha-spending function that has been pre-specified for OS, and then use the critical value that is read from the rejection boundary at the relevant information time. However, this suggestion may arguably be unfair when OS is analyzed only once at the time of trial termination or at best no more frequently than TTP. The question is what should be the best analysis for OS? Please use the method discussed in this chapter to answer this question.

10.3 Study the following case presented by FDA statisticians (Hung, O'Neill, Wang, and Lawrence, 2006): In a heart failure trial, the primary endpoint is a new composite score endpoint of several clinical outcomes (e.g., death, MI, hospitalization for worsening heart failure, change in quality of life status at some time after the patient is treated with a study drug). The sample-size and the total number of composite endpoint events are increased, based on a new projected effect size for the composite endpoint calculated at an interim analysis time of the trial. An important question is how to test each component of the composite endpoints and other clinically important secondary endpoints after this adaptation.

10.4 An adaptive approach for bivariate-endpoint is studied by Todd (2003). Discuss the differences between Todd's method and the method provided in this Chapter.

Chapter 11

Drop-Loser and Add-Arm Design

11.1 Opportunity

An adaptive seamless phase II/III design is one of the most attractive adaptive designs. A seamless adaptive design is a combination of traditional phase II and phase III trials. In seamless design, there is usually a so-called learning phase that serves the same purpose as a traditional phase II trial, followed by a confirmatory phase that serves the same objectives as a traditional phase III trial (Figure 11.1). Compared to traditional designs, a seamless design can reduce sample-size and time-to-market for a positive drug candidate. The main feature of a seamless design is the drop-loser mechanism. Sometimes it also allows for adding new treatment arms. A seamless design usually starts with several arms or treatment groups. At the end of the learning phase, inferior arms (losers) are identified and dropped from the confirmatory phase.

Hung, Wang, O'Neill, and Lawrence from FDA (2006) articulate that it may be advisable to redistribute the remaining planned sample-size of a terminated arm to the remaining treatment arms for comparison so that coupled with use of a proper valid adaptive test, one may enhance the statistical power of the design to detect a dose that is effective (Hung, O'Neill, Wang, and Lawrence, 2006).

In this chapter, we will discuss different methods for seamless designs. Examples are provided to illustrate how to design seamless adaptive trials using the SAS macro.

11.1.1 *Impact Overall Alpha Level and Power*

A seamless design can enjoy the following advantages of potential savings by early stopping for futility and efficacy. A seamless design is efficient because there is no lead time between the learning and confirmatory phases, and

the data collected at the learning phase are combined for final analysis. A noticeable feature of the seamless phase II/III design is that there are differences in controlling type-I error rate (alpha), and power between a seamless design and the traditional design with separate phase II and phase III trials. In traditional designs, if we view the two Phase II and III trials as a super experiment, the actual α is equal to $\alpha_{II}\alpha_{III}$, where α_{II} and α_{III} are the type-I error rates controlled at phase II and phase III, respectively. If two phase III trials are required, then $\alpha = \alpha_{II}\alpha_{III}\alpha_{III}$. In seamless phase II/III design, actual $\alpha = \alpha_{III}$; if two phase III trials are required, then $\alpha = \alpha_{III}\alpha_{III}$. Thus, the α for a seamless design is actually $1/\alpha_{II}$ times larger than the traditional design. Similarly, in the classic "super experiment", the actual power is equal to Power_{II} Power_{III}, while in a seamless phase II/III design, actual power is equal to Power_{III}, where Power_{II} and Power_{III} are the power for phase II and III trials, respectively. Therefore, the power for a seamless design is $1/Power_{II}$ times larger than the traditional design.

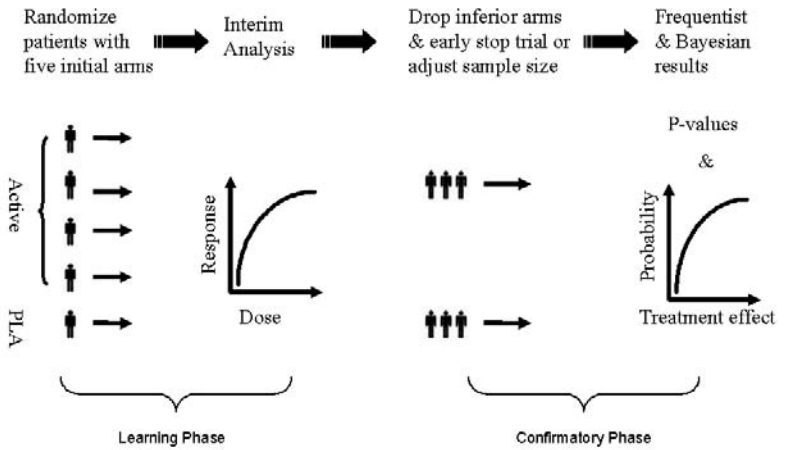


Figure 11.1: Seamless Design

11.1.2 *Reduction In Expected Trial Duration*

As pointed out by Walton (2006), time between studies has multiple components: (1) analysis of observed data, (2) interpretation of analyzed results, (3) planning next study, (4) resource allocation, (5) selection of, agreements with, investigators, (6) IRB submission and approval, and (7) other. In a seamless design, we move the majority of the take "planning next study" to up-front; perform analysis at real time; and combined traditional two IRB submissions and approvals into one. Also, in seamless design there is one set of "selection of, agreements with, investigators" instead of two. Adap-

tive designs require adaptive or dynamic allocation of resources. At the end of traditional phase-IIb design, the analysis and interpretations of results are mainly performed by sponsor and the "go and no-go" decision-making is fully made by sponsor unless there is major safety concern. In seamless design, the traditional phase IIb becomes the first phase of the seamless design and IDMC has a big inference on the decision. From that point of view, seamless design is less biased.

There could be competing constraints among "faster," "cheaper," and "better" as noted by Walton (Figure 11.2). It is challenging to satisfy all goals simultaneously. Decision theory in conjunction with adaptive design is a way to balance the limits to satisfying these goals.

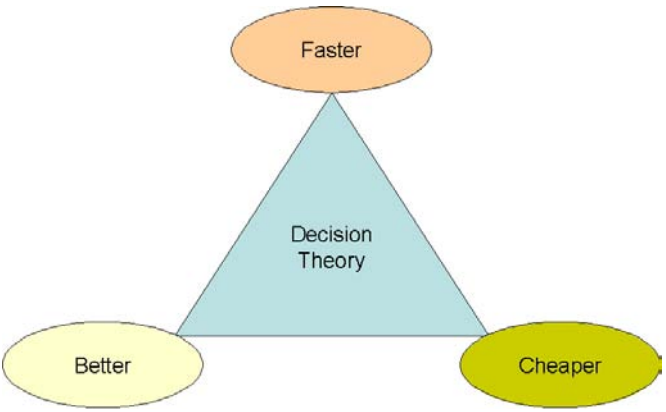


Figure 11.2: Decision Theory for Competing Constraints

11.2 Method with Weak Alpha-Control

11.2.1 *Contract Test Based Method*

The weak $\alpha-$control method for drop-loser design uses a contrast test at the first stage (Chang, 2006). At the interim analysis, the arm with the best observed response and the control will be carried on to the second phase.

In Section 2.4, we have discussed the contrast test:

$$H_o : \sum_{i=1}^{m} c_i u_i - \delta \leq 0 \text{ vs. } H_o : \sum_{i=1}^{m} c_i u_i - \delta > 0. \tag{11.1}$$

The test statistic of a contrast test at the first stage is given by

$$T_1 = \frac{\sum_{i=1}^{m} c_i \hat{u}_i - \delta}{\sigma \sqrt{\sum_{i=1}^{m} \frac{c_i^2}{n_{1i}}}}, \tag{11.2}$$

where $\sigma^2 = var(\hat{u}_i)$ is considered fixed for large sample-size; n_{1i} is the sample-size for the i^{th} arm at the first stage. $\hat{u}_i$ is an estimate of u_i which is the true response for the i^{th} arm. We choose equal contrasts, i.e., $c_1 = m-1$ and $c_i = -1$ $(i = 2, ..., m)$, where m = number of arms and the first arm is the control arm.

The test statistic has a normal distribution:

$$T_1 \sim N(\frac{u_i \sum_{i=1}^{m} c_i - \delta}{\sigma \sqrt{\sum_{i=1}^{m} \frac{c_i^2}{n_{1i}}}}, 1). \tag{11.3}$$

Under the global null hypothesis H_o, the test statistic T_1 has the standard normal distribution. Denote the corresponding p-value as

$$p_1 = 1 - \Phi(T_1). \tag{11.4}$$

Let

$$T_1 = p_1$$

and

$$T_2 = p_1 + p_2,$$

where p_2 is the stagewise p-value based on subsample at stage 2 for the null hypothesis $H_{o2} : u_R = u_1$. Here u_R is the true response (parameter) of the best response arm observed at stage 1, i.e., $\hat{u}_R = \max\{\hat{u}_1, ..., \hat{u}_m\}$ at the first stage, and R is the best observed response arm at stage 1. Without loss of generality, assume $R = m$. Under the global null hypothesis $H_o : u_i = c$, p_1 and p_2 are uniformly distributed over [0,1]. Therefore, the method controls α under the global null hypothesis. If the global null is rejected, we can conclude that there is a treatment difference.

11.2.2 *Sampson-Sill's Method*

Sampson and Sill (2005) proposed a drop-loser design, in which only one winner and the control can be carried on to the second phase of the seamless design. They considered using a conditional distribution on a specific event to construct conditionally unbiased estimators. Following a similar idea, tests and confidence intervals can be derived. The specific conditional distribution used in the final analysis depends on the outcomes from the

first stage. The conditional level α tests are also unconditional level α tests. In other words, if all of the null hypotheses are true for all of the treatments, the probability of rejecting a null hypothesis is never greater than α, regardless of which treatment is selected (Sampson and Sill, 2005).

Sampson and Sill assume the following ordered outcome after the first stage,

$$Q = \{X : \bar{X}_1 > \bar{X}_2 > ... > \bar{X}_m\}$$

so that the first arm τ_1 continues into the second stage (other outcomes would be equivalently handled by relabeling). Typically, at the end of the trial, we want to make inferences on the mean μ_1 of treatment τ_1, and compare it with the control (e.g., test $H_0 : \mu_1 - \mu_0 \leq \Delta_1$ or construct a confidence interval about $\mu_1 - \mu_0$). Therefore, we can construct uniformly most powerful (UMP) unbiased tests for hypotheses concerning Δ_1 based upon the conditional distribution of W. To see this, note that the test is based on the conditional event Q. Lehmann (1983) gives a general proof for this family, which shows that conditioning on sufficient statistics associated with nuisance parameters causes them to drop from the distribution. In addition, the theorem states that the use of this conditional distribution to draw inferences about a parameter of interest yields hypothesis testing procedures which are uniformly most powerful unbiased unconditionally (i.e., UMPU before conditioning on the sufficient statistics). The calculations to carry out the hypothesis and confidence intervals require numerical integration (Sampson and Sill, 2005).

11.2.3 *Normal Approximation Method*

Let's consider a drop-loser design with three arms: A, B, and C (control). At the first stage, we perform two pair-wise comparisons: A versus C and B versus C. The corresponding p-values are denoted by p_{1A} and p_{1B}. If $p_{1A} < p_{1B}$, arm B is dropped; otherwise, arm A is dropped. If we don't drop any arm, the p-value for the two comparisons at the final stage are denoted by p_A and p_B. It is obvious that p_{1A}, p_{1B}, p_A, and p_B are uniformly distributed under the global null that implies all three arms are equally effective. For the drop-loser design, we are interested in the statistic

$$Z^w = I_{AB} Z^A + (1 - I_{AB}) Z^B,$$

where $I_{AB} = 1$ if arm B is dropped; otherwise $I_{AB} = 0$; $Z^A = \Phi(1 - p_A)$ and $Z^B = \Phi(1 - p_B)$.

Shun, Soo, and Lan (2007) found that under the global null hypothesis, Z^w is approximately normal distributed with mean $E(Z^w) = \sqrt{\frac{\tau}{2\pi}}$, and $var(Z^w) = 1 - \frac{\tau}{2\pi}$, where information time for the interim analysis $\tau = \frac{n_1}{n}$ (the sample-size fraction at the interim analysis). Therefore they proposed a test statistic as

$$Z^{ws} = \frac{Z^w - \sqrt{\frac{\tau}{2\pi}}}{\sqrt{1 - \frac{\tau}{2\pi}}},$$

which has approximately the standard normal distribution.

The approximate p-value hence can be easily obtained: $p_A = 1 - \Phi(Z^{ws})$. The exact p-value is given by

$$p = p_A + 0.0003130571\,(4.444461^{\tau}) - 0.00033$$

Note that we have used the inverse-normal transformation to generalize Shen-Soo-Lan method to binary and survival endpoints.

The advantage of this method is simplicity!

11.3 Method with Strong Alpha-Control

11.3.1 *Bauer-Kieser Method*

Since the weak controlled method cannot tell exactly which treatment is effective, Bauer-Kieser (1999) developed a strong control method by adjusting the p-value from the first stage p_1 using single-step method such as Dunnett or Bonferroni method. B-K method is conservative and is OK for design with a larger early efficacy stopping probability. However, for seamless design, the first stage is often for dose selection only. In such cases, B-K method is inefficient. The SAS macro later in this chapter will use the Bonferroni adjusted p-value for the first stage p_1, which is uniformly larger than the p-value with the uniform distribution. In other words, we conservatively choose the test statistic $(M-1)p_1$ for the first stage and $(M-1)p_1 + p_2$ for the second stage, where M is the maximum number of arms in the trial.

11.3.2 *MSP with Single-Step Multiplicity Adjustment*

Assume there are m_1 comparisons among M treatment groups at the first stage. These comparisons can be expressed as m_1 null hypotheses:

$$H_{oi},\ i = 1, ..., m_1. \tag{11.5}$$

The corresponding p-values are $p_{1i},\ i = 1, ...m_1$. With Bonferroni adjustment (if there is a common control group for all the comparisons, Dunnett method is better), the Bonferroni adjusted p-value is $\tilde{p}_{1i} = m_1\, p_{1i}$.

Decision rules are described as follows:

At stage 1: (1) If $m_1\, p_{1i} \leq \alpha_1$, then reject H_{oi} $(i = 1, ..., m_1)$; (2) If $m_1\, p_{1i} > \beta_1$, then accept H_{oi} $(i = 1, ..., m_1)$; (3) If $\beta_1 \geq m_1\, p_{1\,\min} > \alpha_1$, then continue to the second stage and make adaptations (e.g., adjust sample-size and add new arms) if necessary, where $p_{1\,\min} = \min\{p_{11}, ..., p_{1m_1}\}$.

At stage 2: (1) Choose a set of comparisons based on the corresponding p-values $\tilde{p}_{1i}$ or other criteria such as safety, for the second stage. Assume there are m_2 comparisons at the second stage. (2) Based on the second stage data, the naive stagewise p-values are calculated as p_i and the Bonferroni adjusted p-value is $\tilde{p}_{2i} = m_2 p_{2i}$.

Decision rules at stage 2: If $\tilde{p}_{1i} + \tilde{p}_{2i} = m_1 p_{1i} + m_2 p_{2i} \leq \alpha_2$, then reject H_{oi} $(i = 1, ..., m_2)$, otherwise don't reject the null. The global null can be rejected as long as $m_1 p_{1\,\min} + m_2 p_{2\,\min} \leq \alpha_2$, where $\mathrm{p}_{2\,\min} = \{p_{21}, ..., p_{2m_2}\}$.

Alternatively, we can use a fixed sequence test procedure for the second stage. The sequence is determined by the order of the p-values at the first stage (from the smallest p-value to the largest p-value). This procedure is expected to have more power.

11.3.3 *A More Powerful Method*

In the weak control method based on contrast test, the test statistics are $T_1 = p_1$and $T_2 = p_1 + p_2$, respectively, for the first and second stage, where p_2 is the stagewise p-value based on subsample at stage 2 for the null hypothesis $H_{o2} : u_R = u_1$. Assume $R = m$.

To control the type-I error, we need to control α under the condition that the best observed arm is in fact an ineffective arm with the same response as the control arm, i.e.,

$$P\left(T_2 \leq \alpha_2 \cap \beta_1 > T_1 > \alpha_1 | u_1 = u_R\right) \leq \alpha. \tag{11.6}$$

Controlling type-I error under all null confutations are difficult. Let's find out the worst scenario, then we control type-I error under this condition. In other words, we want to find out the null condition that maximizes

the conditional error $p^* = P\left(T_2 \leq \alpha_2 \cap \beta_1 > T_1 > \alpha_1 | u_1 = u_m\right)$. Note that $var(T_1, T_2) = var(T_1, T_1) = 1$. Also, we have

$$\frac{\partial}{\partial u_i} P\left(\min(\alpha_2 - p_2, \beta_1) > T_1 > \alpha_1 | u_1 = u_R\right) = 0, \tag{11.7}$$

where p_2 and $T_1 = p_1$ are conditionally independent under $u_1 = u_m$; $p_2 \sim 1$ and $T_1 \sim f_{u_1 = u_m}(u_1, ..., u_m)$.

There is no simple global maximum solution for (11.7). However, a local maximum solution can be found.

If (11.7) has a unique solution for u_i, then because of exchangeability of $u_2, ..., u_{m-1}$ in (11.6) (p_1^* is a function of $u_1, ..., u_m$), we know immediately that an extreme value occurs when

$$u_2 = u_3 =, ..., = u_{m-1}. \tag{11.8}$$

The other extreme values occur when some of $u_i s$ approach infinity. This will give the minimum value 0 because $\Pr\left(\beta_1 > p_1^* > \alpha_1, u_1 = u_R\right) = 0$.

We now can use simulation to determine the stopping boundaries under condition

$$u_1 = u_m;\ u_2 = u_3 =, ..., = u_{m-1} \tag{11.9}$$

with strong alpha-control.

For a given $u_1 = u_m = u_0$, α_1 and β_1, simulate under various values for $u_2 = u_3 =, ..., = u_{m-1}$ to obtain worst value for $u_2 = u_3 =, ..., = u_{m-1}$, say u^*. Then change α_2 until the percentage of significance is equal to the overall α.

11.4 Application of SAS Macro for Drop-Loser Design

SAS Macro 11.1, DrpLsrNRst, can be used to simulate the trial with drop-loser design using either weak or strong alpha-control. The weak control only controls α under the global null hypothesis. For the strong control, α is controlled under all null configurations. At the first stage, Bonferroni adjustment is used for the strong control by inflating the p-value from p_1 to (nArms-1)$p_{1\,\min}$, where $p_{1\,\min}$ is the smallest p-value for all the comparisons from the first stage. For the weak control, the first stage p-value p_1 is from a contrast test (see Chapter 2) and no p-value adjustment is required. The overall α control is controlled by using MSP. The SAS variables are defined as follows. **nArms** = number of arms in the trial, **us**$\{i\}$ = the true

response (mean, rate, and hazard rate) in the i^{th} arm, **sigma** = common standard deviation, **NId** = noninferiority margin, **N** = sample per group, **Nmax** = maximum sample-size per group, **cPower** = the target conditional power at the interim analysis, **AveN** = average total sample-size, **Alpha1** = early efficacy stopping boundary (one-sided), **beta1**= early futility stopping boundary, **Alpha2** = the final efficacy stopping boundary. For the strong control, **CntlType** = "strong"; otherwise, the weak control is used. The first arm must be the control arm.

>>SAS Macro 11.1: Two-Stage Drop-Loser Adaptive Design>>

```
%Macro DrpLsrNRst(nSims=100000, CntlType="strong", nArms=5,
    alpha=0.025, beta=0.2, NId=0, cPower=0.9, nInterim=50,
    Nmax=150, nAdj="Y", alpha1=0.01, beta1=0.15,
    alpha2=0.1871, EP="normal", sigma=1, tStd=24, tAcr=10);
Data DrpLsrNRst; Set dInput;
 Keep FSP ESP AveN Power cPower Nmax;
 Array us{&nArms}; Array u1{&nArms};
 Array u2{&nArms}; Array cs{&nArms};
 seedx=1736; seedy=6214; alpha=&alpha; NId=&NId;
 Nmax=&Nmax; nArms=&nArms; n1=&nInterim; sigma=&sigma;
 cPower=&cPower; FSP=0; ESP=0; AveN=0; Power=0;
 If &EP="mean" Then sigma=&sigma;
 If &EP="binary" Then sigma=(us1*(1-us{1}))**0.5;
 If &EP="survival" Then
    sigma=us{1}*(1+exp(-us{1}*&tStd)*(1-exp(us{1}*&tAcr))
                 /(&tAcr*us{1}))**(-0.5);

 Do isim=1 to &nSims;
 TotalN=nArms*n1; uMax=us{1}; Cntrst=0; SumSqc=0;
    Do i=1 To nArms;
    u1{i}=Rannor(seedx)*sigma/Sqrt(n1)+us{i};
    If u1{i}>uMax Then Do uMax=u1{i}; iMax=i; End;
    Cntrst=Cntrst+cs{i}*u1{i};
    SumSqc=SumSqc+cs{i}*cs{i};
     End;
    Z1 = Cntrst*Sqrt(n1)/Sqrt(SumSqc)/sigma;
    p1=1-ProbNorm(Z1);
    * Bonferroni Adjustment;
    If &CntlType="strong" Then
      p1=(nArms-1)*(1-ProbNorm((uMax-us{1})/sigma*Sqrt(n1/2)));
    If p1>&beta1 Then FSP=FSP+1/&nSims;
```

```
      If p1<=&alpha1 Then do;
      Power=Power+1/&nSims; ESP=ESP+1/&nSims;
      End;
      If iMax=1 Then Goto myJump;
      If p1>&alpha1 and p1<=&beta1 Then do;
      BF=Probit(1-max(0,&alpha2-p1))-Probit(1-cPower);
      n2=2*(sigma/(u1{iMax}-u1{1})*BF)**2;
      nFinal=min(n1+n2, Nmax);
      If &nAdj="N" Then nFinal=Nmax;
      If nFinal>n1 Then do;
           TotalN=2*(nFinal-n1)+nArms*n1;
           u2{1}=Rannor(seedx)*sigma/Sqrt(nFinal-n1)+us{1};
           u2{iMax}=Rannor(seedy)*sigma/Sqrt(nFinal-n1)+us{iMax};
           T2=(u2{iMax}-u2{1}+NId)*Sqrt(nFinal-n1)/2**0.5/sigma;
      p2=1-ProbNorm(T2); TS2=p1+p2;
           If .<TS2<=&alpha2 Then Power=Power+1/&nSims;
      End;
      End;
myJump:
      AveN=AveN+TotalN/&nSims;
 End;
 Output;
Run;
Proc Print Data=DrpLsrNRst (obs=1); Run;
%Mend DrpLsrNRst;
‹‹SAS‹‹
```

Example 11.1 Seamless Design of Asthma Trial

The objective of this trial in an asthma patient is to confirm sustained treatment effect, measured as FEV1 change from baseline to the 1-year of treatment. Initially, patients are equally randomized to four doses of the new compound and a placebo. Based on early studies, the estimated FEV1 change at week 4 are 6%, 12%, 13%, 14%, and 15% (with pooled standard deviation 18%) for the placebo (dose level 0), dose level 1, 2, 3, and 4, respectively. One interim analysis is planned when 50% of patients have the efficacy assessments. The interim analysis will lead to either picking the winner (arm with best observed response), or stopping trial for efficacy or futility. The winner and placebo will be used at stage 2. The final analysis will be based on the sum of the stagewise p-values from both stages. The stopping boundaries are $\alpha_1 = 0.01$, $\beta_1 = 0.15$, $\alpha_2 = 0.1871$. The decision rules are: if $\tilde{p}_1 \leq \alpha_1$, stop the trial and claim efficacy; if $\tilde{p}_1 > \beta_1$, stop the

trial and claim futility; if $\alpha_1 < \tilde{p}_1 \leq \beta_1$, proceed to the second stage. At the final analysis, if $\tilde{p}_1 + p_2 \leq \alpha_2$, claim efficacy, otherwise claim futility. For the weak control, $\tilde{p}_1 = p_1$, where p_1 is the naive stagewise p-value from a contrast test based on subsample from stage 1. For the strong control, the $\tilde{p}_1$ is the adjusted p-value, i.e., $\tilde{p}_1 = 4p_{1\,\min}$.

Table 11.1: Response and Contracts in Seamless Design

Arms	0	1	2	3	4
FEV1 change	0.06	0.12	0.13	0.14	0.15
Contrasts	-0.54	0.12	0.13	0.14	0.15

The mean changes in FEV1 from baseline and contrast coefficients are presented in Table 11.1. For both the weak-control and strong-control designs, the interim analysis is performed based on 50 patients per group; the maximum sample-size is Nmax = 100; the target conditional power is 90% for sample-size adjustment using MSP.

The simulations are performed for both the null hypothesis (A mean FEV1 change of 6% for all treatment groups) and the alternative hypothesis, using the following SAS macro calls. The results are summarized in Tables 11.2 and 11.3. We can see that the sample-size required for the weak and strong α controls are similar. However, the futility stopping probability is higher for the strong control than the weak control under the null hypothesis; under the alternative, the futility stopping probability is lower for the strong control than the weak control.

>>SAS>>

```
Title "Simulations of Drop-loser Design under Ho for Example 11.1";
Data dInput;
Array us{5} (.06, .06, .06, .06, .06); Array cs{5} (-0.54, .12, .13, .14, .15);
%DrpLsrNRst( CntlType="weak", nArms=5, alpha=0.025, beta=0.2,
  NId=0, cPower=0.9, nInterim=50, Nmax=100, nAdj="Y",
  alpha1=0.01, beta1=0.15, alpha2=0.1871, EP="normal", sigma=0.18);
%DrpLsrNRst(CntlType="strong", nArms=5, alpha=0.025, beta=0.2,
   NId=0, cPower=0.9, nInterim=50, Nmax=100, nAdj="Y",
  alpha1=0.01, beta1=0.15, alpha2=0.1871, EP="normal", sigma=0.18);
Run;

Title "Simulations of Drop-loser Design under Ha for Example 11.1";
Data dInput;
Array us{5} (.06, .12, .13, .14, .15); Array cs{5} (-0.54, .12, .13, .14, .15);
%DrpLsrNRst(CntlType="weak", nArms=5, alpha=0.025, beta=0.2,
  NId=0, cPower=0.9, nInterim=50, Nmax=100, nAdj="Y",
```

```
  alpha1=0.01, beta1=0.15, alpha2=0.1871, EP="normal", sigma=0.18);
%DrpLsrNRst(CntlType="strong", nArms=5, alpha=0.025, beta=0.2,
  NId=0, cPower=0.9, nInterim=50, Nmax=100, nAdj="Y",
  alpha1=0.01, beta1=0.15, alpha2=0.1871, EP="normal", sigma=0.18);
Run;
```

‹‹**SAS**‹‹

Table 11.2: Results of Seamless Design under Global Ho

α-Control	N1	Nmax	FSP	ESP	AveN	cPower	Power
Weak	50	100	.85	.010	264	0.9	.025
Strong	50	100	.97	.002	252	0.9	.002

Table 11.3: Results of Seamless Design Under Ha

α-Control	N1	Nmax	FSP	ESP	AveN	cPower	Power
Weak	50	100	.053	.634	278	0.9	0.896
Strong	50	100	.009	.552	285	0.9	0.903

11.5 Summary and Discussion

We have studied the seamless designs that allow for dropping losers and/or adding new arms, and early efficacy or futility stopping. Note that the efficiency of a seamless design is sensitive to the sample-size fraction or information time in the end of the learning phase. Therefore simulations should be done to determine the best information time for the interim analysis.

Practically, the seamless trials require early efficacy readouts. This early efficacy assessment can be the primary endpoint for the trial or surrogate endpoint (biomarker). Because data analysis and interpretation allow exploration of richness of clinical data, the interim analysis should also include some variables other than the primary. Those analyses can be descriptive or hypothesis testing kind. The inferastruct such CDISC, EDC, ExpDesign Studio®, and East®, Clinical Data Workbench® can be helpful in this regard. Seamless design can be also used for other situations such as a combination of Phase I and Phase II trials. Regarding the logistic issues in a seamless design, please see the papers by PhRMA adaptive working group (Maca, et al., 2006; Quinlan, Gallo, Krams, 2006).

Problem

11.1 Study the effect of contrasts (Table 11.4) on the power and other operating characteristics using a five-group drop-loser design with and without sample-size adjustment.

Table 11.4: Response and Contrast Shapes

Shape	u_0	u_1	u_2	u_3	u_4	c_0	c_1	c_2	c_3	c_4
Monotonic	1.0	2.0	3.5	4.0	4.5	-1.9	-0.9	0.1	1.1	1.6
Wave	1.0	1.0	4.0	1.0	3.0	-1.0	-1.0	2.0	-1.0	1.0
Step	1.0	3.4	3.4	3.4	3.4	-1.92	0.48	0.48	0.48	0.48

11.2 Study drop-loser designs with a binary endpoint.

11.3 Conduct a numerical study of the weak $\alpha-$control method using SAS Macro 11.1 under different null configurations:
(1) $\delta_i = c$ (2) $\delta_1 = \delta_2 = c$, $\delta_i > c$. (3)....

11.4 Suppose in a classic m-group dose-finding trial with a continuous response (variance $\sigma^2 = 1$), the test statistics for pairwise comparisons are defined as

$$Z_i = \frac{\bar{x}_i - \bar{x}_0}{\sqrt{n}},$$

where $\bar{x}_i$ is the mean of the i^{th} group, $i = 0$ for the control group, and n is the sample-size per group. Study the distributions of the following order statistics:

$$Z_{(1)} <, ..., < Z_{(m)}.$$

Are the distributions close to the distributions for the standard order statistics (Kokoska and Zwillianer, 2000)?

11.5 Implement SAS macros or R functions for drop-loser designs with MPP and MINP.

Chapter 12

Biomarker-Adaptive Design

12.1 Opportunities

Biomarkers, as compared to a true endpoint such as survival, can often be measured earlier, easier, and more frequently; are less subject to competing risks, and less confounded. The utilization of biomarker will lead to a better target population with a larger effect size, a smaller sample-size required, and faster decision-making. With the advancement of proteomic, genomic and genetic technologies, personalized medicine with the right drug for the right patient becomes possible.

Conley and Taube (2004) described the future of biomarker/genomic markers in cancer therapy: "The elucidation of the human genome and fifty years of biological studies have laid the groundwork for a more informed method for treating cancer with the prospect of realizing improved survival. Advanced in knowledge about the molecular abnormalities, signaling pathways, influence the local tissue milieu and the relevance of genetic polymorphism offer hope of designing effective therapies tailored for a given cancer in particular individual, as well as the possibility of avoiding unnecessary toxicity."

Wang, Hung, and O'Neill (2006) from FDA have pointed out: "Generally, when the primary clinical efficacy outcome in a phase III trial requires much longer time to observe, a surrogate endpoint thought to be strongly associated with the clinical endpoint may be chosen as the primary efficacy variable in phase II trials. The results of the phase II studies then provide an estimated effect size on the surrogate endpoint, which is supposedly able to help size the phase III trial for the primary clinical efficacy endpoint, where often it is thought to have a smaller effect size."

What exactly is a biomarker? National Institutes of Health Workshop (Gruttola, 2001) gave the following definitions. *Biomarker* is a characteristic that is objectively measured and evaluated as an indicator of normal

biologic processes, pathogenic processes, or pharmacological responses to a therapeutic intervention. *Clinical endpoint* (or outcome) is a characteristic or variable that reflects how a patient feels or functions, or how long a patient survives. *Surrogate endpoint* is a biomarker intended to substitute for a clinical endpoint. Biomarkers can also be classified as classifier, prognostic, and predictive biomarkers.

A classifier biomarker is a marker, e.g., a DNA marker, that usually does not change over the course of study. A classifier biomarker can be used to select the most appropriate target population or even for personalized treatment. For example, a study drug is expected to have effects on a population with a biomarker, which is only 20% of the overall patient population. Because the sponsor suspects that the drug may not work for the overall patient population, it may be efficient and ethical to run a trial only for the subpopulations with the biomarker rather than the general patient population. On the other hand, some biomarkers such as RNA markers are expected to change over the course of the study. This type of markers can be either a prognostic or predictive marker.

A prognostic biomarker informs the clinical outcomes, independent of treatment. It provides information about natural course of the disease in individual with or without treatment under study. A prognostic marker does not inform the effect of the treatment. For example, NSCLC patients receiving either EGFR inhibitors or chemotherapy have better outcomes with a mutation than without a mutation. Prognostic markers can be used to separate good and poor prognosis patients at the time of diagnosis. If expression of the marker clearly separates patients with an excellent prognosis from those with a poor prognosis, then the marker can be used to aid the decision about how aggressive the therapy needs to be. The poor prognosis patients might be considered for clinical trials of novel therapies that will, hopefully, be more effective (Conley and Taube, 2004). Prognostic markers may also inform the possible mechanisms responsible for the poor prognosis, thus leading to the identification of new targets for treatment and new effective therapeutics.

A predictive biomarker informs the treatment effect on the clinical endpoint. A predictive marker can be population-specific: a marker can be predictive for population A but not population B. A predictive biomarker, as compared to true endpoints like survival, can often be measured earlier, easier, and more frequently and is less subject to competing risks. For example, in a trial of a cholesterol-lowering drug, the ideal endpoint may be death or development of coronary artery disease (CAD). However, such a study usually requires thousands of patients and many years to conduct. Therefore, it is desirable to have a biomarker, such as a reduction in post-

treatment cholesterol, if it predicts the reductions in the incidence of CAD. Another example would be an oncology study where the ultimate endpoint is death. However, when a patient has disease progression, the physician will switch the patient's initial treatment to an alternative treatment. Such treatment modalities will jeopardize the assessment of treatment effect on survival because the treatment switching is response-adaptive rather than random (See Chapters 13 and 14). If a marker, such as time-to-progression (TTP) or response rate (RR), is used as the primary endpoint, then we will have much cleaner efficacy assessments because the biomarker assessment is performed before the treatment switching occurs.

In this chapter, we will discuss adaptive designs using classifier, prognosis, and predictive markers. The challenges in marker validations will also be discussed.

12.2 Design with Classifier Biomarker

12.2.1 *Setting the Scene*

As mentioned earlier, a drug might have different effects in different patient populations. A hypothetical case is presented in Table 12.1, where RR_+ and RR_- are the response rates for biomarker-positive and biomarker-negative populations, respectively. In the example, there is a treatment effect of 25% in the 10 million patient population with the biomarker, but only 9% in the 50 million general patient population. The sponsor faces the dilemma of whether to target the general patient population or use biomarkers to select a smaller set of patients that are expected to have a bigger response to the drug.

Table 12.1: Response Rate and Sample-Size Required

	Population	RR_+	RR_-	Sample-size
Biomarker (+)	10M	50%	25%	160*
Biomarker (-)	40M	30%	25%	
Total	50M	34%	25%	1800

Note: *800 subjects screened. Power = 80%.

There are several challenges: (1) The estimated effect size for each subpopulation at the design stage is often very inaccurate; (2) A cost is associated with screening patients for the biomarker; (3) The test for detecting the biomarker often requires a high sensitivity and specificity, and the screening tool may not be available at the time of the clinical trial; (4) Screening patients for the biomarker may cause a burden and impact

patient recruitment. These factors must be considered in the design.

Ideally, the utility function should be constructed first in order to decide which population we should target for. There are many utility functions to choose. For example, the utility can be defined as $U = \Sigma\left(\delta_i N_i - C_i\right)$, where δ_i is the effect size of the i^{th} subpopulation with the size of N_i and C_i is the associated cost or loss.

Suppose we decide to run a trial on population with a biomarker. It is interesting to study how the screening testing impact the expected utility. The size of the target patient size N with biomarker (+) can be expressed as

$$N = N_+ S_e + N_-(1 - S_p), \tag{12.1}$$

where N_+ and N_- are the sizes of patient populations with and without the biomarker, respectively; S_e is the sensitivity of the screening test, i.e., the probability of correctly identifying the biomarker among patients with the biomarker, and S_p is the specificity of the screening test, which is defined as the probability of correctly identifying biomarker-negative among patients without biomarker. The average treatment effect for diagnostic biomarker (+) patients:

$$\Delta = \frac{\Delta_+ N_+ S_e + \Delta_- N_-(1 - S_p)}{N}. \tag{12.2}$$

If the utility is defined as the overall benefit for the patient population screened as biomarker-positive, i.e., $U = \Delta N$, then the expected utility is given by

$$U_e = \Delta N \; Power. \tag{12.3}$$

Figure 12.1 shows how the specificity will impact the target population size, the average treatment effect in the target population, and the expected utility under different designs. When the specificity increases, the target population decreases, but the average treatment effect in the target population increases because misdiagnosis of biomarker-negative as positive will reduce the average treatment effect.

Using adaptive design, we can start with the overall patient population. At the interim analysis, a decision can be made whether to go for the subpopulation or the overall population based on the expected utilities:

(1) If we target for the subpopulation with the biomarker, the expected utility at interim analysis is given by

(conditional power of subpopulation) $\times$ *(Impact of success)*
- (1-conditional power of subpopulation) $\times$ *(Impact of failure)*

(2) If we target for the full patient population, the expected utility at interim analysis is given by

(conditional power of full population) x (Impact of success)
- (1-conditional power of full population) x (Impact of failure)

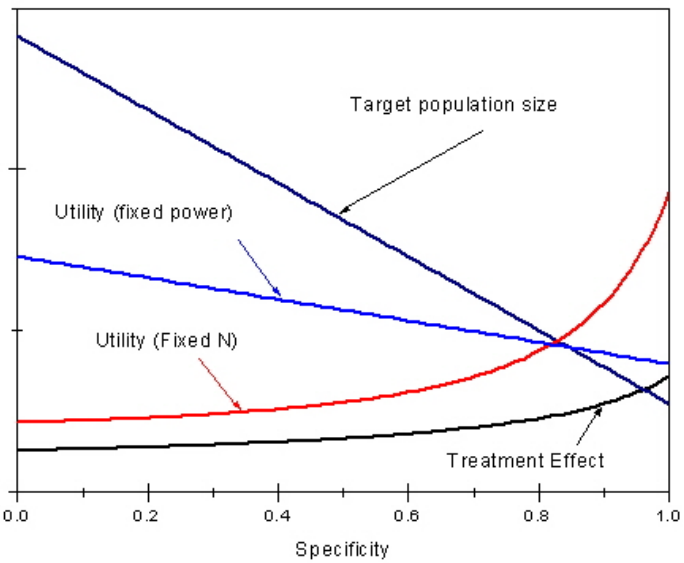


Figure 12.1: Effect of Biomarker Misclassification

12.2.2 *Classic Design with Classifier Biomarker*

Denote treatment difference between the test and control groups by δ_+, δ_-, and δ, for biomarker-positive, biomarker-negative, and overall patient populations, respectively. The null hypothesis for biomarker-positive subpopulation is

$$H_{o1} : \delta_+ = 0. \tag{12.4}$$

The null hypothesis for biomarker-negative subpopulation is

$$H_{o2} : \delta_- = 0. \tag{12.5}$$

The null hypothesis for overall population is

$$H_o : \delta = 0. \tag{12.6}$$

Without loss of generality, assume that the first n patients have the biomarker among N patients and the test statistic for the subpopulation is given by

$$Z_+ = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i}{n\sigma}\sqrt{\frac{n}{2}} \sim N(0,1) \text{ under } H_o, \tag{12.7}$$

where x_i, and y_i $(i = 1, ..., n)$ are the responses in treatment A and B.

Similarly, the test statistic for biomarker-negative group is defined as

$$Z_- = \frac{\left(\sum_{i=n+1}^{N} x_i - \sum_{i=n+1}^{N} y_i\right)}{(N-n)\,\sigma}\sqrt{\frac{N-n}{2}} \sim N(0,1) \text{ under } H_o. \tag{12.8}$$

The test statistic for overall population is given by

$$Z = \frac{\hat{\delta}}{\sigma}\sqrt{\frac{N}{2}} = T_+\sqrt{\frac{n}{N}} + T_-\sqrt{\frac{N-n}{N}} \sim N(0,1) \text{ under } H_o. \tag{12.9}$$

We choose the test statistic for the trial as

$$T = \max(Z, Z_+). \tag{12.10}$$

It can be shown that the correlation coefficient between Z and Z_+ is

$$\rho = \sqrt{\frac{n}{N}}. \tag{12.11}$$

Therefore, the stopping boundary can be determined by

$$\Pr(T \geq z_{2,\,1-\alpha} | H_o) = \alpha, \tag{12.12}$$

where $z_{2,\,1-\alpha}$ is the bivariate normal $100(1-\alpha)$-equipercentage point under H_o.

The p-value corresponding to an observed test statistic t is given by

$$p = \Pr\left(T \geq t | H_o\right). \tag{12.13}$$

The power can be calculated using

$$\Pr\left(T \geq z_{2,\,1-\alpha} | H_a\right) = \alpha. \tag{12.14}$$

The numerical integration or simulations can be performed to evaluate $z_{2,\,1-\alpha}$ and the power.

Note that the test statistic for the overall population can be defined as

$$Z = w_1 Z_+ + w_2 Z_-,$$

where w_1 and w_2 are constants satisfying $w_1^2 + w_2^2 = 1$. In such case, the correlation coefficient between Z and Z_+ is $\rho = w_1$.

More generally, if there are m groups under consideration, we can define a statistic for the g^{th} group as

$$Z_g = \frac{\hat{\delta}_g}{\sigma}\sqrt{\frac{n_g}{2}} \;\sim\; N(0,1) \text{ under } H_o. \tag{12.15}$$

The test statistic for the overall population is given by

$$T = \max\left\{Z_1, ..., Z_g\right\}, \tag{12.16}$$

where $\{Z_1, ..., Z_m\}$ is asymptotically m-variate standard normal distribution under H_o with expectation $\mathbf{0} = \{0, ..., 0\}$ and correlation matrix $\mathbf{R}=\{\rho_{ij}\}$. It can be easily shown that the correlation between Z_i and Z_j is given by

$$\rho_{ij} = \sqrt{\frac{n_{ij}}{n_i n_j}}, \tag{12.17}$$

where n_{ij} is the number of concordant pairs between the i^{th} and j^{th} groups.

The asymptotic formulation for power calculation with the multiple tests is similar to that for multiple-contrast tests (Bretz and Hothorn, 2002):

$$\begin{aligned}
&\Pr\left(T \geq z_{m,\,1-\alpha} | H_a\right) \\
&= 1 - \Pr(Z_1 < z_{m,\,1-\alpha} \;\cap\; ... \;\cap\; T_m < z_{m,\,1-\alpha} \;|H_a \\
&= 1 - \Phi_m\left(\left(\mathbf{z}_{m,\,1-\alpha} - e\right) diag\left(\frac{1}{v_0}, ..., \frac{1}{v_m}\right);\; \mathbf{0};\; \mathbf{R}\right),
\end{aligned}$$

where $\mathbf{z}_{m,\,1-\alpha} = (z_{m,\,1-\alpha}, ..., z_{m,\,1-\alpha})$ stands for the m-variate normal $100(1-\alpha)$-equipercentage point under H_o, $\mathbf{e} = (E_a(T_0), ..., E_a(T_m))$ and $\mathbf{v} = (v_0, ..., v_m) = \left(\sqrt{V_0\,(T_0)}, \sqrt{V_1\,(T_1)}, ..., \sqrt{V_1\,(T_m)}\right)$ are vectorially summarized expectations and standard errors.

The power is given by

$$p = \Pr\left(T \geq z_{m,\,1-p}\right). \tag{12.18}$$

For other types of endpoints, we can use inverse-normal method, i.e., $Z_g = \Phi\,(1-p_g)$ in (12.15), where p_g is the p-value for the hypothesis test in the g^{th} population group, then (12.17) and (12.18) are still approximately valid.

Simulation Algorithm

(1) Generate n_+ responses for biomarker-positive population (BPP).
(2) Generate n_- responses for biomarker-negative population (BNP).
(3) Compute test statistic T_+ for BPP and T_o for overall population.
(4) Compute $T = \max\{T_+, T_o\}$.
(5) Repeat (1)-(4) many times and compute the percentage of the outcomes with $T > Z_c$. This percentage is probability $\Pr\left(T > Z_c\right)$.

To determine the critical point Z_c for rejecting the null at α level, run the simulations under the null condition for various Z_c until $\Pr\left(T > Z_c\right) \approx \alpha$. To determine the power, run the simulations under the alternative condition, the power is given by $\Pr\left(T > Z_c\right)$ or the percentage of the outcomes with $T > Z_c$.

12.2.3 *Adaptive Design with Classifier Biomarker*

Strong Alpha-Controlled Method

Let the hypothesis test for biomarker-positive subpopulation at the first stage (size = n_1/group) be

$$H_{o1} : \delta_+ \leq 0 \tag{12.19}$$

and the hypothesis test for overall population (size = N_1/group) be

$$H_o : \delta \leq 0 \tag{12.20}$$

with the corresponding stagewise p-values, p_{1+} and p_1, respectively. These stagewise p-values should be adjusted. A conservative way is used Bonferroni method or a method similar to Dunnett method that takes the

correlation into consideration. For Bonferroni-adjusted p-value and MSP, the test statistic is $T_1 = 2\min(p_{1+}, p_1)$ for the first stage. The population with a smaller p-value will be chosen for the second stage and the test statistic for the second stage is defined as $T_2 = T_1 + p_2$, where p_2 is the stagewise p-value from the second stage. This method is implemented in SAS Macro 12.1 as described below.

SAS Macro 12.1 is developed for simulating biomarker-adaptive trials with two parallel groups. The key SAS variables are defined as follows: **Alpha1** = early efficacy stopping boundary (one-sided), **beta1** = early futility stopping boundary, **Alpha2** = final efficacy stopping boundary, **u0p** = response difference in biomarker-positive population, **u0n** = response in biomarker-negative population, **sigma** = asymptotic standard deviation for the response difference, assuming homogeneous variance among groups. For binary response, sigma $=\sqrt{r_1(1-r_1)+r_2(1-r_2)}$; For Normal response, sigma $=\sqrt{2}\sigma$. **np1**, **np2** = sample sizes per group for the first and second stage for the biomarker-positive population. **nn1**, **nn2** = sample sizes per group for the first and second stage for the biomarker-negative population. **cntlType** = "strong," for the strong type-I error control and **cntlType** = "weak," for the weak type-I error control, **AveN** = average total sample-size (all arms combined), **pPower** = the probability of significance for biomarker-positive population, **oPower** = the probability of significance for overall population.

››SAS Macro 12.1: Biomarker-Adaptive Design››

```
%Macro BMAD(nSims=100000, cntlType="strong", nStages=2,
     u0p=0.2, u0n=0.1, sigma=1, np1=50, np2=50, nn1=100,
     nn2=100, alpha1=0.01, beta1=0.15,alpha2=0.1871);
Data BMAD;
 Keep FSP ESP Power AveN pPower oPower;
 seedx=1736; seedy=6214; u0p=&u0p; u0n=&u0n; np1=&np1;
     np2=&np2; nn1=&nn1; nn2=&nn2; sigma=&sigma;
     FSP=0; ESP=0;Power=0; AveN=0; pPower=0; oPower=0;
 Do isim=1 to &nSims;
      up1=Rannor(seedx)*sigma/Sqrt(np1)+u0p;
      un1=Rannor(seedy)*sigma/Sqrt(nn1)+u0n;
      uo1=(up1*np1+un1*nn1)/(np1+nn1);
      Tp1=up1*np1**0.5/sigma; To1=uo1*(np1+nn1)**0.5/sigma;
      T1=Max(Tp1,To1); p1=1-ProbNorm(T1);
      If &cntlType="strong" Then p1=2*p1; *Bonferroni;
      If p1>&beta1 Then FSP=FSP+1/&nSims;
      If p1<=&alpha1 Then Do;
```

```
            Power=Power+1/&nSims; ESP=ESP+1/&nSims;
            If Tp1>To1 Then pPower=pPower+1/&nSims;
            If Tp1<=To1 Then oPower=oPower+1/&nSims;
        End;
        AveN=AveN+2*(np1+nn1)/&nSims;
        If &nStages=2 And p1>&alpha1 And p1<=&beta1 Then Do;
            up2=Rannor(seedx)*sigma/Sqrt(np2)+u0p;
            un2=Rannor(seedy)*sigma/Sqrt(nn2)+u0n;
            uo2=(up2*np2+un2*nn2)/(np2+nn2);
            Tp2=up2*np2**0.5/sigma;  To2=uo2*(np2+nn2)**0.5/sigma;
            If Tp1>To1 Then Do;
              T2=Tp2; AveN=AveN+2*np2/&nSims;
           End;
            If Tp1<=To1 Then Do;
              T2=To2; AveN=AveN+2*(np2+nn2)/&nSims;
          End;
            p2=1-ProbNorm(T2); Ts=p1+p2;
            If .<TS<=&alpha2 Then Do;
              Power=Power+1/&nSims;
              If Tp1>To1 Then pPower=pPower+1/&nSims;
              If Tp1<=To1 Then oPower=oPower+1/&nSims;
            End;
        End;
    End;
Run;
Proc Print Data=BMAD (obs=1); Run;
%Mend BMAD;
```

‹‹SAS‹‹

Example 12.1 Biomarker-Adaptive Design

Suppose in an active-control trial, the estimated treatment difference is 0.2 for the biomarker-positive population (BPP) and 0.1 for the biomarker-negative population (BNP) with a common standard deviation of $\sigma = 1$. Using SAS Macro 12.1, we can generate the operating characteristics under the global null hypothesis H_o (u0p = 0, u0n = 0), the null configurations H_{o1} (u0p = 0, u0n = 0.1) and H_{o2} (u0p = 0.2, u0n = 0), and the alternative hypothesis H_a (u0p = 0.2, u0n = 0.1) (See Table 12.2). Typical SAS macro calls to simulate the global null and the alternative conditions are presented in the following.

```
>>SAS>>
Title "Simulation under global Ho, 2-stage design";
%BMAD(nSims=100000, CntlType="strong", nStages=2, u0p=0,
    u0n=0, sigma=1.414, np1=260, np2=260, nn1=520, nn2=520,
    alpha1=0.01, beta1=0.15,alpha2=0.1871);

Title "Simulations under Ha, 2-stage design";
%BMAD(nSims=100000, CntlType="strong", nStages=2, u0p=0.2,
    u0n=0.1, sigma=1.414, np1=260, np2=260, nn1=520, nn2=520,
    alpha1=0.01, beta1=0.15,alpha2=0.1871);
<<SAS<<
```

To generate the corresponding results for the classic single stage design (See Table 12.3 for the simulation results), we can use the SAS calls as follows:

```
>>SAS>>
Title "Simulations under global Ho, single-stage design";
%BMAD(nSims=100000, CntlType="strong", nStages=1, u0p=0,
u0n=0, sigma=1.414, np1=400, np2=0, nn1=800, nn2=0, alpha1=0.025);

Title "Simulations under Ha, single-stage design";
%BMAD(nSims=100000, CntlType="strong", nStages=1, u0p=0.2,
u0n=0.1, sigma=1.414, np1=400, np2=0, nn1=800, nn2=0, alpha1=0.025);
<<SAS<<
```

Table 12.2: Simulation Results of Two-Stage Design

Case	FSP	ESP	Power	AveN	pPower	oPower
H_o	0.876	0.009	0.022	1678	0.011	0.011
H_{o1}	0.538	0.105	0.295	2098	0.004	0.291
H_{o2}	0.171	0.406	0.754	1852	0.674	0.080
H_a	0.064	0.615	0.908	1934	0.311	0.598

H_{o1} and H_{o2} = no effect for BPP and overall population.

Table 12.3: Simulation Results of Classic Single-Stage Design

Case	FSP	ESP	Power	AveN	pPower	oPower
H_o	0.878	0.022	0.022	2400	0.011	0.011
H_{o1}	0.416	0.274	0.274	2400	0.003	0.271
H_{o2}	0.070	0.741	0.741	2400	0.684	0.056
H_a	0.015	0.904	0.904	2400	0.281	0.623

H_{o1} and H_{o2} = no effect for BPP and overall population.

Trial monitoring is particularly important for these types of trials. Assume we have decided the sample sizes N_2 per treatment group for overall population at stage 2, of which n_2 (can be modified later) subjects per group are biomarker-positive. Ideally, decision on whether the trial continues for the biomarker-positive patients or overall patients should be dependent on the expected utility at the interim analysis. The utility is the total gain (usually as a function of observed treatment effect) subtracted by the cost due to continuing the trial using BPP or the overall patient population. For simplicity, we define the utility as the conditional power. The population group with larger conditional power will be used for the second stage of the trial. Suppose we design a trial with $n_{1+} = 260$, $n_{1-} = 520$, $p_{1+} = 0.1$, $p_1 = 0.12$, and stopping boundaries: $\alpha_1 = 0.01, \beta_1 = 0.15$, and $\alpha_2 = 0.1871$. For $n_{2+} = 260$, and $n_{2-} = 520$, the conditional power based on MSP is 82.17% for BPP and 99.39% for the overall population. The calculations are presented as follows:

$$P_c\left(p_1, \delta\right) = 1 - \Phi\left(\Phi^{-1}\left(1 - \alpha_2 + p_1\right) - \frac{\delta}{\sigma}\sqrt{\frac{n_2}{2}}\right), \alpha_1 < p_1 \leq \beta_1.$$

For the biomarker-positive population,

$$\Phi^{-1}\left(1 - 0.1871 + 0.1\right) = \Phi^{-1}(0.912\,9) = 1.3588, 0.2\sqrt{260/2} = 2.280\,4,$$

$$P_c = 1 - \Phi\left(1.3588 - 2.2804\right) = 1 - \Phi(-0.921\,6) = 1 - 0.1783 = 0.821\,7.$$

For the biomarker-negative population,

$$\Phi^{-1}\left(1 - 0.1871 + 0.12\right) = \Phi^{-1}(0.932\,9) = 1.4977,$$

$$0.2\sqrt{(260 + 520)/2} = 3.949\,7,$$

$$P_c = 1 - \Phi\left(1.4977 - 3.949\,7\right) = 1 - \Phi(-2.452) = 1 - 0.0071 = 0.992\,9.$$

Therefore, we are interested in the overall population. Of course, different n_2 and N_2 can be chosen at the interim analyses, which may lead to different decisions regarding the population for the second stage.

The following aspects should also be considered during design: power versus utility, enrolled patients versus screened patients, screening cost, and the prevalence of biomarker.

12.3 Challenges in Biomarker Validation

12.3.1 *Classic Design with Biomarker Primary-Endpoint*

Given the characteristics of biomarkers, can we use a biomarker as the primary endpoint for late-stage or confirmatory trials? Let's study the outcome in three different scenarios. (1) The treatment has no effect on the true endpoint or the biomarker. (2) The treatment has no effect on the true endpoint but does affect the biomarker. (3) The treatment has a small effect on the true endpoint but has a larger effect on the biomarker. Table 12.4 summarizes the type-I error rates (α) and powers for using the true endpoint and biomarker under different scenarios. In the first scenario, we can use either the true endpoint or biomarker as the primary endpoint because both control the type-I error. In the second scenario, we cannot use the biomarker as the primary endpoint because α will be inflated to 81%. In the third scenario, it is better to use the biomarker as the primary endpoint from a power perspective. However, before the biomarker is fully validated, we don't know which scenario is true; use of the biomarker as the primary endpoint could lead to a dramatic inflation of the type-I error. It must be validated before a biomarker can be used as primary endpoint.

Table 12.4: Issues with Biomarker Primary Endpoint

Effect size ratio	Endpoint	Power (alpha)
0.0/0.0	True endpoint	(0.025)
	Biomarker	(0.025)
0.0/0.4	True endpoint	(0.025)
	Biomarker	(0.810)
0.2/0.4	True endpoint	0.300
	Biomarker	0.810

Note: N = 100 per group. Effect size ratio = effect size of true endpoint to effect size of biomarker.

12.3.2 *Treatment-Biomarker-Endpoint Relationship*

Validation of biomarker is not an easy task. Validation here refers to the proof of a biomarker to be a predictive marker, i.e., a marker can be used as a surrogate marker. Before we discuss biomarker validations, let's take a close look at the 3-way relationships among treatment, biomarker, and the true endpoint. It is important to be aware that the correlations between them are not transitive. In the following example, we will show that it could be the case that there is a correlation (R_{TB}) between treatment and the biomarker and a correlation (R_{BE}) between the biomarker and the true

endpoint, but there is no correlation (R_{TE}) between treatment and the true endpoint (Figures 12.2 and 12.3).

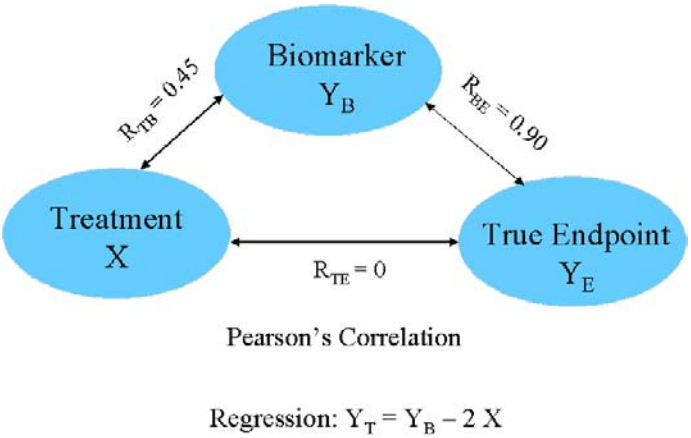


Figure 12.2: Treatment-Biomarker-Endpoint Three-Way Relationship

The hypothetical example to be discussed is a trial with 14 patients, 7 in the control group and 7 in the test group. The biomarker and true endpoint outcomes are displayed in Figure 12.3. The results show that the Pearson's correlation between the biomarker and the true endpoint is 1 (perfect correlation) in both treatment groups. If the data are pooled from the two groups, the correlation between the biomarker and the true endpoint is still high, about 0.9. The average response with the true endpoint is 4 for each group, which indicates that the drug is ineffective compared with the control. On the other hand, the average biomarker response is 6 for the test group and 4 for the control group, which indicates that the drug has effects on the biomarker.

Facing the data, what we typically do is to fit a regression model with the data, in which the dependent variable is the true endpoint (Y_T) and the independent variables (predictors) are the biomarker (Y_B) and the treatment (X). After model fitting, we can obtain that

$$Y_T = Y_B - 2X. \tag{12.21}$$

This model fits the data well based on model-fitting p-value and R^2. Specifically, R^2 is equal to 1, p-values for model and all parameters are equal to 0, where the coefficient 2 in model (12.21) is the separation between the two lines. Based on (12.21), we would conclude that both biomarker and treatment affect the true endpoint. However, we know that the treatment has no effect on biomarker at all.

In fact, the biomarker predicts the response in the true endpoint, but it does not predict the treatment effect on the true endpoint, i.e., it is a prognostic marker.

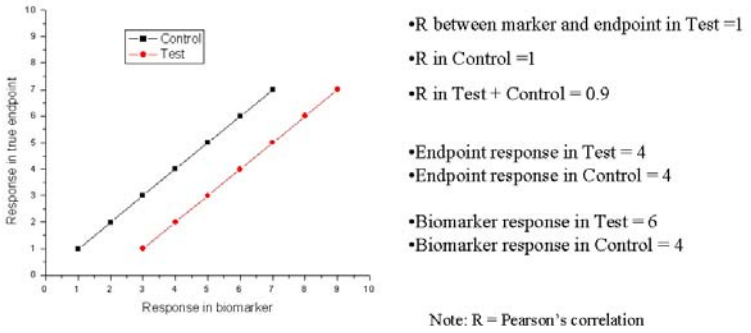


Figure 12.3: Correlation Versus Prediction

12.3.3 *Multiplicity and False Positive Rate*

Let's further discuss the challenges from a multiplicity point of view. In earlier phases or the discovery phase, we often have a large number of biomarkers to test. Running hypothesis testing on many markers can be done either with a high false positive rate without multiplicity adjustment or a low power with multiplicity adjustment. Also, if model selection procedures are used without multiplicity adjustment as we commonly see in current practice, the false positive rate could be inflated dramatically. Another source of false positive discovery rate is the so-called publication bias. The last, but not least, source of false positive finding is due to the multiple testing conducted by different companies or research units. Imagine that 100 companies study the same biomarker, even if family-wise type-I error rate is strictly controlled at a 5% level within each company, there will still be, on average, 5 companies that have positive findings about the same biomarker just by chance.

12.3.4 *Validation of Biomarkers*

We now realized the importance of biomarker validation and would like to review some commonly used statistical methods for biomarker validation.

Prentice (1989) proposed four operational criteria: (1) treatment has a significant impact on the surrogate endpoint; (2) treatment has a significant impact on the true endpoint; (3) the surrogate has a significant impact on the true endpoint; and (4) the full effect of treatment upon the true

endpoint is captured by the surrogate endpoint. Note that this method is for a binary surrogate (Molenberghs, et al., 2005).

Freedman, et, al. (1992) argued that the last Prentice criterion is difficult statistically because it requires that the treatment effect is not statistically significant after adjustment of the surrogate marker. They further articulated that the criterion might be useful to reject a poor surrogate marker, but it is inadequate to validate a good surrogate marker. Therefore they proposed a different approach based on the proportion of treatment effect on true endpoint explained by biomarkers and a large proportion required for a good marker. However, as noticed by Freedman, this method is practically infeasible due to the low precision of the estimation of the proportion explained by the surrogate.

Buyse and Molenberghs (1998) proposed the internal validation matrices, which include relative effect (RE) and adjusted association (AA). The former is a measure of association between the surrogate and the true endpoint at an individual level, and the latter expresses the relationship between the treatment effects on the surrogate and the true endpoint at a trial level. The practical use of the Buyse-Molenberghs method raises a few concerns: (1) a wide confidence interval of RE requires a large sample-size; (2) treatment effects on the surrogate and the true endpoint are multiplicative, which cannot be checked using data from a single trial.

Other methods, such as external validation using meta-analysis and two-stage validation for fast track programs, also face similar challenges in practice. For further readings on biomarker evaluations, Weir and Walley (2006) give an excellent review; Case and Qu (2006) proposed a method for quantifying the indirect treatment effect via surrogate markers and Alonso et al. (2006) proposed a unifying approach for surrogate marker validation based on Prentice's criteria.

12.3.5 *Biomarkers in Reality*

In reality, there are many possible scenarios: (1) same effective size for the biomarker and true endpoint, but the biomarker response can be measured earlier; (2) bigger effective size for the biomarker and smaller for the true endpoint; (3) no treatment effect on the true endpoint, limited treatment effect on the biomarker; and (4) treatment effect on the true endpoint only occurs after the biomarker response reaches a threshold. Validation of biomarkers is challenging, and the sample-size is often insufficient for the full validation. Therefore validations are often performed to a certain degree and soft validation scientifically (e.g., pathway) is important.

What is the utility of partially validated biomarkers? In the next sec-

tion, we will discuss how to use prognostic markers in adaptive designs.

12.4 Adaptive Design with Prognostic Biomarker

12.4.1 *Optimal Design*

A biomarker before it is proved predictive can only be considered as a prognostic marker. In the following example, we discuss how to use a prognostic biomarker (a marker may be predictive) in trial design. The adaptive design proposed permits early stopping for futility based on the interim analysis of the biomarker. At the final analysis, the true endpoint will be used to preserve the type-I error. Assume there are three possible scenarios: (1) H_{o1}: effect size ratio ESR = 0/0, H_{o2}: effect size ratio ESR = 0/0.25, and (3) H_a : effect size ratio ESR = 0.5/0.5, but biomarker response earlier. ESR is the ratio of effect size for true endpoint to the effect size for biomarker. We are going to compare three different designs: classic design and two adaptive designs with different stopping boundaries as shown in Table 12.5.

Table 12.5: Adaptive Design with Biomarker

Design	Condition	Power	Expected N/arm	Futility boundary
	H_{o1}		100	
Classic	H_{o2}		100	
	H_a	0.94	100	
	H_{o1}		75	
Adaptive	H_{o2}		95	β_1= 0.5
	H_a	0.94	100	
	H_{o1}		55	
Adaptive	H_{o2}		75	β_1= 0.1056
	H_a	0.85	95	

Based on simulation results (Table 12.5), we can see that the two adaptive designs reduce sample-size required under the null hypothesis. However, this comparison is not good enough because it does not consider the prior distribution of each scenario at the design stage.

We have noticed that there are many different scenarios with associated probabilities (prior distribution) and many possible adaptive designs with associated probabilistic outcomes (good and bad). Suppose we have also

formed the utility function, the criteria for evaluating different designs. Now let's illustrate how we can use utility theory to select the best design under financial, time, and other constraints.

Table 12.6: Prior Knowledge About Effect Size

Scenario	Effect Size Ratio	Prior Probability
H_{o1}	0/0	0.2
H_{o2}	0/0.25	0.2
H_a	0.5/0.5	0.6

Let's assume the prior probability for each of the scenarios mentioned earlier as shown in Table 12.6. For each scenario, we conduct computer simulations to calculate the probability of success and the expected utilities for each design. The results are summarized in Table 12.7.

Table 12.7: Expected Utilities of Different Designs

Design	Classic	Biomarker-adaptive	
		$\beta_1 = 0.5$	$\beta_1 = 0.1056$
Expected Utility	419	441	411

Based on the expected utility, the adaptive design with the stopping boundary $\beta_1 = 0.5$ is the best. Of course, we can also generate more designs and calculate the expected utility for each design and select the best one.

12.4.2 *Prognostic Biomarker in Designing Survival Trial*

Insufficiently validated biomarker such as tumor response rate (RR) can be used in oncology trial for interim decision-making whether to continue to enroll patients or not to reduce the cost. When the response rate in the test group is lower, because of the correlation between RR and survival, it is reasonable to believe the test drug will be unlikely to have survival benefit. However, even when the trial stopped earlier due to unfavorable results in response rate, the survival benefit can still be tested. We have discussed this for a Non-Hodgkin's Lymphoma (NHL) trial in Chapter 10.

12.5 Adaptive Design with Predictive Marker

If a biomarker is proved to be predictive, then we can use it to replace the true-endpoint from the hypothesis test point of view. In other words, a proof of treatment effect on predictive marker is a proof of treatment effect on the true endpoint. However, the correlation between the effect sizes of treatment in the predictive (surrogate) marker and the true endpoints is desirable but unknown. This is one of the reasons that follow-up study on the true endpoint is highly desirable in the NDA accelerated approval program.

Changes in biomarker over time can be viewed as stochastic process (marker process) and have been used in the so-called threshold regression (Chapter 13). A predictive marker process can be viewed an external process that covariates with the parent process. It can be used in tracking progress of the parent process if the parent process is latent or is only infrequently observed. In this way, the marker process forms a basis for predictive inference about the status of the parent process of clinical endpoint. The basic analytical framework for a marker process conceives of a bivariate stochastic process $\{X(t), Y(t)\}$ where the parent process $\{X(t)\}$ is one component process and the marker process $\{Y(t)\}$ is the other. Whitmore, Crowder, and Lawless (1998) investigated the failure inference based on a bivariate. Wiener model has also been used in this aspect, in which failure is governed by the first-hitting time of a latent degradation process. Lee, DeGruttola, and Schoenfeld (2000) apply this bivariate marker model to CD4 cell counts in the context of AIDS survival. Hommel, Lindig, and Faldum (2005) studied a two-stage adaptive design with correlated data.

12.6 Summary and Discussion

We have discussed the adaptive designs with classifier, prognostic, and predictive markers. These designs can be used to improve the efficiency by identifying the right population, making decisions earlier to reduce the impact of failure and delivering the efficacious and safer drugs to market earlier. However, full validation of a biomarker is statistically challenging and sufficient validation tools are not available. Fortunately, adaptive designs with biomarkers can be beneficial even when the biomarkers are not fully validated. The Bayesian approach is an ideal solution for finding an optimal design, while computer simulation is a powerful tool for the utilization of biomarkers in trial design.

Problem

12.1 Develop SAS macros or R function for the method proposed in Section 12.2.2 and conduct a simulation study of the method.

Chapter 13

Adaptive Treatment Switching and Crossover

13.1 Treatment Switching and Crossover

To study the efficacy of a test drug for progressive disease such as cancer or HIV, a parallel-group active-control randomized clinical trial is often used. Under the study design, qualified patients are randomly assigned to receive either an active control or the test treatment under investigation. Patients are allowed to switch from one treatment to another due to ethical considerations such as lack of response or if there is evidence of disease progression. In practice, it is not uncommon that up to 80% of patients may switch from one treatment to another. This certainly has an impact on the evaluation of the efficacy of the test treatment. Despite the switching between different treatments or interventions, many clinical studies are to compare the test drug with the active control agent as if no patients had ever switched.

The effect of treatment switching will impact the survival curve. The patterns of the curves before and after switching are expected to be different, for example, they may follow a mixed exponential distribution with different hazard rates before and after switching.

Sommer and Zeger (1991) referred to the treatment effect among patients who complied with treatment as *biological efficacy*. The survival time of a patient who switched from the active control to the test treatment might be on the average longer than his/her survival time that would have been if he/she had adhered to the original treatment, if switching is based on prognosis to optimally assign patients' treatments over time. We refer to the difference caused by treatment switch as *switching effect*.

The purpose of this chapter is to provide useful models for modeling clinical trials with response-adaptive treatment switching. The remaining of this chapter is organized as follows. A mixed exponential model is considered to assess the total survival time in Section 13.2. A mixture of Wiener

processes is studied in Section 13.3. In Section 13.4, the concept of latent hazard rate is considered by incorporating the switching effect in the latent hazard functions. Summary and discussions will be presented in Section 13.5.

13.2 Mixed Exponential Survival Model

In clinical trials, the target patient population often consists of two or more subgroups based on heterogeneous baseline characteristics (e.g., the patients could be a mixture of the second-line and the third-line oncology patients). The median survival time of the third-line patients is usually shorter than that of the second-line patients. If the survival times of the two subgroup populations are modeled by exponential distributions with hazard rates λ_1 and λ_2, respectively, then the survival distribution of the total population in the trial is a mixed exponential distribution with a probability density function of

$$P_1\lambda_1 e^{-\lambda_1 t} + P_2\lambda_2 e^{-\lambda_2 t}(t > 0),$$

where t is the survival time and P_1 and P_2 (fixed or random) are the proportions of the two sub-populations. Following the similar idea of Mendenhal and Hader (1985), the maximum likelihood estimates of the parameters λ_i and P_i can be obtained. Also, if time to each disease progression is exponential distribution, the total survival time is a mixed exponential distribution (Chang, 2005; Chow and Chang, 2006). We will discuss this application in detail.

13.2.1 *Mixed Exponential Model*

In cancer trials, there are often signs/symptoms (or more generally biomarkers) that indicate the state of the disease and the ineffectiveness or failure of a treatment. A cancer patient often experiences several episodes of progressed diseases before death. Therefore, it is natural to construct a survival model based on the disease mechanism. In what follows, we consider a mixed exponential model, which is derived from the more general mixed Gamma model.

Let τ_i be the time from the $(i-1)^{th}$ disease progression to the i^{th} disease progression, where $i = 1, .., n$. τ_i is assumed to be mutually independent with probability density function of $f_i(\tau_i)$. The survival time t for a subject can be written as follows

$$t = \sum_{i=1}^{n} \tau_i. \tag{13.1}$$

Note that the n^{th} disease progression is death. The following lemma regarding the distribution of linear combination of two random variables is useful.

Lemma The probability density function of $z = ax + by$ is given by

$$f_z(z) = \frac{1}{a} \int_{-\infty}^{\infty} f(\frac{z - by}{a}, y)dy, \tag{13.2}$$

where $x \sim \ f_x(x)$ and $y \sim \ f_y(y)$.

Proof.

$$F_z(z) = P(Z \leq z) = \iint\limits_{ax+by \leq z} f(x, y)dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\frac{z-by}{a}} f(x, y)dxdy.$$

Take the derivative with respect to z and exchange the order of the two limit processes, (13.2) is immediately obtained. □

Corollary When x and y are independent, then

$$f_z(z) = \frac{1}{a} \int_{-\infty}^{\infty} f_x(\frac{z - by}{a}) f_Y(y)dy. \tag{13.3}$$

Theorem 13.1 If n independent random variables $\tau_i, \ \ i = 1, ..., n$ are exponentially distributed with parameter λ_i, i.e.,

$$\tau_i \sim \ \ f_i(\tau_i) = \lambda_i e^{-\lambda_i \tau_i}, \ (\tau_i \geq 0),$$

then the probability density function of random variable $t = \sum_{i=1}^{n} \tau_i$ is a mixed Gamma distribution, given by

$$f(t; n) = \sum_{i=1}^{n} \frac{\lambda_i e^{-\lambda_i t}}{\prod_{\substack{k=1 \\ k \neq i}}^{n} (1 - \frac{\lambda_i}{\lambda_k})}, \ t > 0, \tag{13.4}$$

where $\lambda_i \neq \lambda_k$ if $k \neq i$ for $i, \ k \in m_0 \leq n$ and m_i is the number of replicates for λ_i with the same value.

Proof. By mathematical induction, when n = 2, Lemma (13.2) gives ($\lambda_i \neq \lambda_k$ if $i \neq k$)

$$f(t;2) = \lambda_1\lambda_2 \int_0^t \exp(-\lambda_1 t - (\lambda_2 - \lambda_1)\tau_2)d\tau_2 = \frac{\lambda_1 e^{-\lambda_1 t}}{1 - \frac{\lambda_1}{\lambda_2}} + \frac{\lambda_2 e^{-\lambda_2 t}}{1 - \frac{\lambda_2}{\lambda_1}}.$$

Therefore, (13.4) is proved for $n = 2$.

Now assume (13.4) hold for any $n \geq 2$, and it will be proven that (13.4) also hold for $n + 1$. From (13.4) and corollary (13.3), we have

$$\begin{aligned} f(t;n+1) &= \int_0^t f(t - \tau_{n+1}; n) f_{n+1}(\tau_{n+1}) d\tau_{n+1} \\ &= \int_0^t \sum_{i=1}^{n} \frac{\lambda_i e^{-\lambda_i (t-\tau_{n-1})}}{\prod_{\substack{k=1 \\ k\neq i}}^{n} (1 - \frac{\lambda_i}{\lambda_k})} \lambda_{n+1} e^{-\lambda_{n+1}\tau_{n+1}} d\tau_{n+1} \\ &= \sum_{i=1}^{n} \frac{1}{\prod_{\substack{k=1 \\ k\neq i}}^{n} (1 - \frac{\lambda_i}{\lambda_k})} \left[\frac{\lambda_i e^{-\lambda_i t}}{1 - \frac{\lambda_i}{\lambda_{n+1}}} + \frac{\lambda_{n+1} e^{-\lambda_{n+1} t}}{1 - \frac{\lambda_{n+1}}{\lambda_i}} \right] \\ &= \sum_{i=1}^{n+1} \frac{\lambda_i e^{-\lambda_i t}}{\prod_{\substack{k=1 \\ k\neq i}}^{n+1} (1 - \frac{\lambda_i}{\lambda_k})}. \end{aligned}$$

□

For disease progression, it is usually true that $\lambda_i > \lambda_k$ for $i > k$. Note, however, that $f(t;n)$ does not depend on the order of λ_i in the sequence, and

$$f(t;n)_{\lambda_n \to +\infty} = f(t;n-1).$$

The survival function $S(t)$ can be easily obtained from (13.4) by integration and the survival function is given by

$$S(t;n) = \sum_{i=1}^{n} w_i e^{-\lambda_i t};\ t > 0,\ n \geq 1, \tag{13.5}$$

where the weight is given by

$$w_i = \left[\prod_{k=1, k\neq i}^{n} (1 - \frac{\lambda_i}{\lambda_k}) \right]^{-1}. \tag{13.6}$$

The mean survival time and its variance are given by

$$\mu = \sum_{i=1}^{n} \frac{w_i}{\lambda_i} \quad \text{and} \quad \sigma^2 = \sum_{i=1}^{n} \frac{w_i}{\lambda_i^2},$$

respectively. When $n = 1$, $w_1 = 1$, (13.4) reduces to the p.d.f. of the exponential distribution. It can be shown that the weights have the properties of $\sum_{i=1}^{n} w_i = 1$ and $\sum_{i=1}^{n} w_i \lambda_i = 0$.

13.2.2 *Effect of Patient Enrollment Rate*

We are going to examine the effect of the accrual duration on the survival distribution. Let N be the number of patients enrolled and let $(0, t_0)$ be the patient enrollment period defined as the time elapsed from the first patient enrolled to the last patient enrolled. Also, let t denote the time elapsed from the beginning of the trial. Denote $f_d(t)$ and $f_e(\tau_e)$, where $\tau_e \epsilon [0,\ T_0]$, the probability density function of failure (death) and the patient enrollment rate, respectively. The failure function (or the probability of death before time t) can be expressed as

$$F(t) = \int_0^t f_d(\tau) d\tau = \int_0^t \int_0^{\min(\tau, t_0)} f(\tau - \tau_e) f_e(\tau_e) d\tau_e d\tau. \tag{13.7}$$

For a uniform enrollment rate,

$$f_e(\tau_e) = \begin{cases} \frac{N}{t_0} & \text{if } \tau_e \epsilon [0,\ t_0] \\ 0, & \text{otherwise;} \end{cases}$$

and probability density function (13.4), (13.7) becomes

$$F(t) = \int_0^t \int_0^{\min(\tau, t_0)} \sum_{i=1}^{n} w_i \frac{\lambda_i e^{-\lambda_i (\tau - \tau_e)}}{t_0} d\tau_e d\tau.$$

Carrying out the integration, we have

$$F(t) = \begin{cases} \frac{1}{t_0} \{ t + \sum_{i=1}^{n} \frac{w_i}{\lambda_i} \left[e^{-\lambda_i t} - 1 \right] \} & \text{if } t \le t_0 \\ \frac{1}{t_0} \{ t_0 + \sum_{i=1}^{n} \frac{w_i}{\lambda_i} \left[e^{-\lambda_i t} - e^{-\lambda_i (t - t_0)} \right] & \text{if } t > t_0 \end{cases}. \tag{13.8}$$

The p.d.f. can be obtained by differentiating $F(t)$ with respect to t:

$$f(t) = \begin{cases} \frac{1}{t_0} \left(1 - \sum_{i=1}^{n} w_i e^{-\lambda_i t} \right) & \text{if } t \le t_0 \\ \frac{1}{t_0} \sum_{i=1}^{n} w_i \left[e^{-\lambda_i (t - t_0)} - e^{-\lambda_i t} \right] & \text{if } t > t_0 \end{cases}. \tag{13.9}$$

The survival function is then given by

$$S(t) = 1 - F(t) \tag{13.10}$$

and the number of deaths among N patients can be written as

$$D(t) = NF(t). \tag{13.11}$$

Note that (13.8) is useful for sample-size calculation with a nonparametric method. For $n = 1$, (13.11) reduces to the number of deaths for the exponential survival distribution, i.e.,

$$D = \begin{cases} R(t - \frac{1}{\lambda}e^{-\lambda t}) & \text{if } t \le t_0 \\ R\left[t_0 - \frac{1}{\lambda}(e^{\lambda t_0} - 1)e^{-\lambda t}\right] & \text{if } t > t_0 \end{cases}, \tag{13.12}$$

where the uniform enrollment rate $R = \frac{N}{t_0}$.

Parameter estimate

It is convenient to use the paired variable $\left(\hat{t}_j, \delta_j\right)$ defined as $\left(\hat{t}_j, 1\right)$ for a failure time $\hat{t}_j$ and $\left(\hat{t}_j, 0\right)$ for a censored time $\hat{t}_j$. The likelihood then can be expressed as

$$L = \prod_{j=1}^{N} \left[f(\hat{t}_j)\right]^{\delta_j} \left[S(\hat{t}_j)\right]^{1-\delta_j}, \tag{13.13}$$

where the probability density function $f(t)$ and survival function $S(t)$ are given by (13.4) and (13.5), respectively, for instantaneous enrollment, and (13.9) and (13.10) for uniform enrollment. Note that for an individual whose survival time is censored at $\hat{t}_j$, the contribution to the likelihood is given by the probability of surviving beyond that point in time, i.e., $S(\hat{t}_j)$. To reduce the number of parameters in the model, we can assume that the hazard rates take the form of a geometric sequence, i.e., $\lambda_i = a\lambda_{i-1}$ or $\lambda_i = a^i\lambda_0$; $i = 1, 2, ..., n$. This leads to a two-parameter model regardless of n, the number of progressions. The maximum likelihood estimates of λ and a can be easily obtained through numerical iterations.

Example 13.1 Adaptive Treatment Switching Trial

To illustrate the mixed exponential model for obtaining the maximum likelihood estimates with two parameters of λ_1 and λ_2, independent x_{1j} and x_{2j}, $j = 1, .., N$ from two exponential distributions with λ_1 and λ_2, respectively are simulated. Let $\tau_j = x_{1j} + x_{2j}$. Then τ_j has a mixed exponential distribution with parameters λ_1 and λ_2. Let $\hat{t}_j = \min(\tau_j, T_s)$, where T_s is the duration of the study. Now, we have the paired variables $\left(\hat{t}_j, \delta_j\right)$, $j = 1, ..., N$, which are used to obtain the maximum likelihood estimators

$\hat{\lambda}_1$ and $\hat{\lambda}_2$. Using (13.13) and the invariance principle of maximum likelihood estimators, the maximum likelihood estimate of mean survival time, $\hat{\mu}$, can be obtained as

$$\hat{\mu} = \sum_{j=1}^{2} \frac{\hat{w}_j}{\hat{\lambda}_j} = \frac{1}{\hat{\lambda}_1} + \frac{1}{\hat{\lambda}_2}. \tag{13.14}$$

For each of the three scenarios (i.e., $\lambda_1 = 1$, $\lambda_2 = 1.5$; $\lambda_1 = 1$, $\lambda_2 = 2$; $\lambda_1 = 1$, $\lambda_2 = 5$), $5,000$ simulation runs are performed. The resulting means and coefficients of variation of the estimated parameters are summarized in Table 13.1. As it can be seen from Table 13.1, the mixed exponential model performs well, which gives an excellent estimate of mean survival time for all three cases with a less than 10% coefficient of variation. The maximum likelihood estimate of λ_1 is reasonably good with a bias less than 6%. However, there are about 5% to 15% over-estimates for λ_2 with large coefficients of variation ranging from 30% to 40%. The bias increases as the percent of the censored observations increases. Thus, it is suggested that the maximum likelihood estimate of mean survival time rather than the maximum likelihood estimate of the hazard rate be used to assess the effect of a test treatment.

Table 13.1: Simulation Results with Mixed Exponential Model

	λ_1	λ_2	μ	λ_1	λ_2	μ	λ_1	λ_2	μ
True	1.00	1.50	1.67	1.00	2.00	1.50	1.00	5.00	1.20
Mean*	1.00	1.70	1.67	1.06	2.14	1.51	1.06	5.28	1.20
CV*	0.18	0.30	0.08	0.20	0.37	0.08	0.18	0.44	0.09
PDs		93%			96%			96%	
Censors		12%			8%			5%	

Note: Study duration $T = 3.2$, fast enrollment. Number of subjects =100.
*Mean and CV of the estimates from 5000 runs for each scenario

13.2.3 *Hypothesis Test and Power Analysis*

In a two-arm clinical trial comparing treatment difference in survival, the hypotheses can be written as

$$\begin{aligned} H_0 &: \mu_1 \geq \mu_2 \\ H_a &: \mu_1 < \mu_2. \end{aligned} \tag{13.15}$$

Note that hazard rates for the two treatment groups may change over time. The proportional hazard rates do not generally hold for a mixed exponential model. Nonparametric methods such as the log-rank test (Marubini

and Valsecchi, 1995) are useful. Note that procedure for sample-size calculation using the log-rank test under the assumption of an exponential distribution is available in the literature (see, e.g., Marubini and Valsecchi, 1995; Chang and Chow, 2005). Here we will derive a formula for sample-size calculation under the mixed exponential distribution based on log-rank statistic. The total number of deaths required for a one-sided log-rank test for the treatment difference between two equal-sized independent groups is given by

$$D = \left[z_{1-\alpha} + 2z_{1-\beta} \frac{\sqrt{\theta}}{1+\theta} \right]^2 \left(\frac{1+\theta}{1-\theta} \right)^2, \tag{13.16}$$

where the hazard ratio is

$$\theta = \frac{\ln F_1(T_s)}{\ln F_2(T_s)}. \tag{13.17}$$

T_s is trial duration, and $F_k(T_s)$ is the proportion of patients with the event in the k^{th} group. The relationship between $F_k(T_s)$, t_0, T_s, and hazard rates λ_i is given by (13.8). From (13.8) and (13.16), the total number of patients required for a uniform enrollment can be obtained as follows

$$N = \frac{\left[z_{1-\alpha} + 2z_{1-\beta} \frac{\sqrt{\theta}}{1+\theta} \right]^2 \left(\frac{1+\theta}{1-\theta} \right)^2}{F_1 + F_2}, \tag{13.18}$$

where N = sample-size per group and t_0 is the duration of enrollment.

Example 13.2 Mixed Exponential Model

Assume a uniform enrollment with a duration of $T_0 = 10$ months and the trial duration $T_s = 14$ months. Further assume that at the end of the study, the expected proportions of failures are $F_1 = 0.8$ and $F_2 = 0.75$ for the control (group 1) and the active drug (group 2), respectively. Choose a power of 90% and one-sided $\alpha = 0.025$. The hazard ratio $\theta = 1.29$ is calculated using (13.17). A total sample size of $N = 714$ is obtained from (13.18). If hazard rates are given instead of the proportions of failures, we may use (13.8) to calculate the proportion of failures first.

13.3 Threshold Regression

In Chapter 5, we have discussed Lan-DeMets method with Brownian motion or Wiener process. In this section, we will see a very different use of the stochastic process.

Survival time often can be characterized by the first failure or first hitting time (FHT) of a stochastic process. The individual experiences a clinical endpoint such as death when the corresponding process reaches an adverse threshold state for the first time. The time scale can be calendar time or some other such as information time. The process can be latent or unobservable such as in the case of competing risks. Threshold regression refers to first hitting time models with regression structures that accommodate covariate data (Lee and Whitmore, 2004). The parameters of the process, threshold state, and time scale are usually depend on the covariates.

13.3.1 *First Hitting Time Model*

FHT model has two essential components: (1) a parent stochastic process $\{X(t),\ t \in T\ ,\ x \in X\}$ with initial value $X(0) = x_0$, where T is the time space and X is the state space of the process and (2) a boundary or threshold B, where $B \subset X$. Note that B can be a function of time t or a stochastic process.

Taking the initial value $X(0) = x_0$ of the process to lie outside boundary set B, the first hitting time of B is the random variable S defined as follows: $S = inf\{t : X(t) \in B\}$. Thus, FHT is the time when the stochastic process first reaches the boundary B. We can see that FHT is the stopping time for the adaptive design using Wiener process in Chapter 5. Another example is the Bernoulli process with negative binomial first-hitting-time. The number of trials S required to reach the m^{th} success in a Bernoulli process $\{B_t, t = 1, 2, ...\}$ has a negative binomial distribution with parameters m and p, where p is the success probability on each trial.

Most survival data are gathered under conditions of competing risks in which two or more causes are competing to determine the observed duration such as in the case of treatment switching. Also a death may be caused by multiple medical conditions that compete to produce death (Kalbfleisch and Prentice, 2002; and Crowder, 2001).

FHT models accommodate the competing risk aspect in a natural fashion. The observed FHT time is the smallest latent FHT. The concept of a latent FHT offers an interesting vehicle for discussing competing risks. Latent FHTs, other than the smallest, are generally unobservable.

The parent process $\{X(t)\}$ and boundary B of the FHT model will both generally have parameters that depend on covariates. For example, the Wiener process has mean parameter μ and variance parameter σ^2 and the boundary B has parameter x_0, the initial position. In threshold regression, these parameters will be connected to linear combinations of covariates using suitable regression link functions:

$$g_\theta(\theta_i) = \mathbf{z}_i\boldsymbol{\beta}, \tag{13.19}$$

where $g_\theta(\cdot)$ is the link function, parameter θ_i is the value of parameter for the i^{th} individual, $z_i = (1, z_{i1}, ..., z_{ik})$ is the covariate vector of the i^{th} individual (with a leading unit to include an intercept term) and $\boldsymbol{\beta}$ is the associated vector of regression coefficients. Research work that has considered regression structures for FHT models include Whitmore (1983), Whitmore, Crowder, and Lawless (1998), and Lee, DeGruttola, and Schoenfeld (2000).

13.3.2 *Mixture of Wiener Processes*

Lee, Chang, and Whitmore (2007) propose mixture of Wiener process for clinical trials with adaptive switching and applied to an oncology study, which are outlined as follows.

Running Time

We have already noted that the actual or censored survival time is composed of two intervals, representing the time on the primary therapy and the time on the alternative therapy. In recognition of the fact that the disease may progress at different rates in these two intervals (irrespective of the treatment), we transform survival times from calendar time to the so-called running time which has the following form:

$$r = a_1\tau_1 + \tau_2, \tag{13.20}$$

where τ_1 and τ_2 correspond to time-to-progression and progression-to-death, respectively, and a_1 is a scale parameter to be estimated.

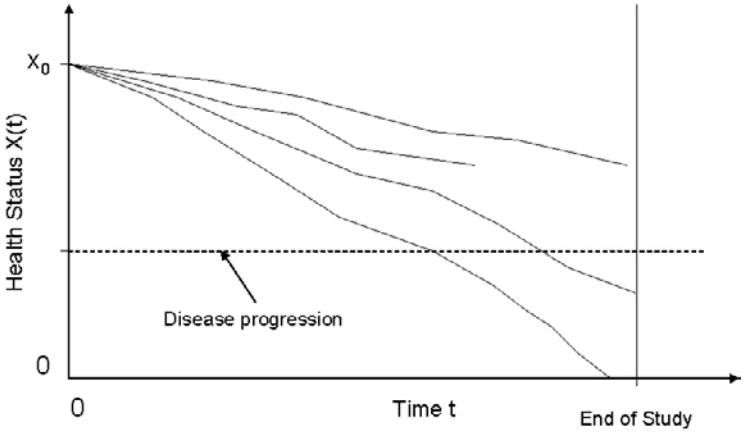


Figure 13.1: Different Paths of Mixed Wiener Process

First Hitting Model

As noted earlier, the first hitting time distribution of a boundary by a Wiener diffusion model, as illustrated in Figure 13.1 for the cancer trial, follows an inverse Gaussian survival distribution. This distribution depends on the initial health status level (x_0) and the mean and variance parameters (μ and σ^2) of the underlying Wiener process. Let $f(r|\mu, \sigma^2, x_0)$ and $F(r|\mu, \sigma^2, x_0)$ denote the probability density function (p.d.f.) and cumulative distribution function (c.d.f.) of the FHT distribution, both defined in terms of running time r. The p.d.f. for the first hitting time is given by

$$f\left(r|\mu,\sigma^2,x_0\right) = \frac{x_0}{\sqrt{2\pi\sigma^2 r^3}} \exp\left[-\frac{\left(x_0+\mu r\right)^2}{2\sigma^2 r}\right], \text{ for } x_0 > 0. \qquad (13.21)$$

If $\mu > 0$, then the FHT is not certain to occur and the p.d.f. is improper. Specifically, in this case, $P(S = \infty) = 1 - exp(-2x_0\mu/\sigma^2)$. The c.d.f. corresponding to (13.21) is given by

$$F\left(r|\mu,\sigma^2,x_0\right) = \Phi\left[-\frac{(\mu r + x_0)}{\sqrt{\sigma^2 r}}\right] + \exp\left[-\frac{2x_0\mu}{\sigma^2}\right]\Phi\left[\frac{\mu r - x_0}{\sqrt{\sigma^2 r}}\right], \text{ for } x_0 > 0, \qquad (13.22)$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution.

Mixture of Wiener Processes

The mixture model for survival time S therefore has the following form:

$$Pr(S > r) = \bar{G}(r) = p\bar{F}_1(r) + (1-p)\bar{F}_2(r). \tag{13.23}$$

The $\bar{F}_j(r)$, $j = 1, 2$, are the respective component survival functions of the mixture and are expressed in terms of running time as defined in equation (13.20). When the parameter p is predetermined based on certain population characteristics (e.g., patients with certain genetic marker), p is the proportion of the patients with the characteristics. Otherwise, p doesn't have an easy interpretation.

Statistical Inference

Each component FHT distribution in mixture model (13.23) is an inverse Gaussian distribution. The three parameters of the FHT distribution are not estimable from survival data because the health status process is latent here. Hence, one parameter may be fixed in each component distribution of the mixture. We set the variance parameter σ^2 to unity.

Each component survival function of the mixture model (13.23) has its own p.d.f. $f_j(r)$ and c.d.f. $F_j(r)$, as well as its own initial health status x_{0j} and mean parameter μ_j , for $j = 1, 2$. For compactness, we now denote the vector of parameters of our mixture model by $\boldsymbol{\theta}$. This vector includes all of the regression coefficients for parameters p, μ_1, x_{01}, μ_2, x_{02}, and a_1, where the last one is the rate parameter in the running time formula (13.20). We let r_i denote the running time for patient i. Time r_i is the running time at death for a dying patient and a right censored running time for death for a surviving patient. Hence, each dying patient i contributes probability density $g(r_i|\boldsymbol{\theta})$ to the sample likelihood function, for $i = 1, ..., n_1$, where $g(r|\boldsymbol{\theta})$ is the mixture p.d.f. corresponding to model (13.26) and n_1 is the number of patients who die before the end of the study. In addition, each surviving patient i contributes survival probability $\bar{G}(r_i|\boldsymbol{\theta}) = 1 - G(r_i|\boldsymbol{\theta})$ to the sample likelihood function, for $i = n_1+1, ..., n_1+n_0$, where $G(r|\boldsymbol{\theta})$ is the mixture c.d.f. in (13.23) and n_0 is the number of patients who survive to the end of the study. We assume that this censoring is noninformative. The sum $n = n_1 + n_0$ is the total number of patients. The sample log-likelihood function to be maximized therefore has the form:

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n_1} \ln g(r_i|\boldsymbol{\theta}) + \sum_{i=n_1+1}^{n_1+n_0} \ln \bar{G}(r_i|\boldsymbol{\theta}).$$

We can use a numerical gradient optimization routine in Stata® to find the maximum likelihood estimate of the regression coefficient vector $\boldsymbol{\theta}$.

13.4 Latent Event Time Model for Treatment Crossover

Suppose that patients are randomly assigned to two treatment groups: a test treatment and an active control. Consider the case where there is no treatment switch and the study objective is to compare the efficacy of the two treatments. Let $T_1, .., T_n$ be independent non-negative survival times and $C_1, ..., C_n$ be independent non-negative censoring times that are independent of survival times. Thus, the observations are $Y_i = min(T_i, C_i)$, $i = 1$ if $T_i \leq C_i$, and $i = 0$ if $T_i > C_i$. Assume that the test treatment acts multiplicatively on a patient's survival time, i.e. an accelerated failure time model applies. Denote the magnitude of this multiplicative effect by $e^{-\beta}$, where β is an unknown parameter.

Consider the situation where patients may switch their treatments and the study objective is to compare the biological efficacy. Let $S_i > 0$ denote the i^{th} patient's switching time. Branson and Whitehead (2002) introduced the concept of *latent event time* in the simple case where only patients in the control group may switch. Here we define the latent event time in the general case as follows. For a patient with no treatment switch, the latent event time is the same as his/her survival time. For patient i who switches at time S_i, the latent event time $\tilde{T}_i$ is an abstract quantity defined to be the patient's survival time that would have been if this patient had not switched the treatment. For patients who switch from the active control group to the test treatment group, Branson and Whitehead (2002) suggested the following model conditional on S_i:

$$\tilde{T}_i \stackrel{d}{=} S_i + e^{\beta} (T_i - S_i), \tag{13.24}$$

where d denotes equality in distribution. That is, the survival time for a patient who switched from the active control to the test treatment could be back-transformed to the survival time that would have been if the patient had not switched.

Branson and Whitehead (2002) proposed an iterative parameter estimation (IPE) method for statistical analysis of data with treatment switch. The idea of the method is to relate the distributions of the survival times of the two treatments under a parametric model. Thus, under model (13.24), IPE can be described as follows. First, an initial estimate $\hat{\beta}$ of β is obtained. Then, latent event times are estimated as

$$\hat{T}_i = S_i + b^{\hat{\beta}} (T_i - S_i) \tag{13.25}$$

for patients who switched their treatments. Next, a new estimate of β

is obtained by using the estimated latent event times as if they were the observed data. Finally, the previously described procedure is iterated until the estimate of β converges.

For the case where patients may switch from either group, model (13.24) can be modified as follows

$$\tilde{T}_i \overset{d}{=} S_i + e^{\beta(1-2k_i)} (T_i - S_i), \tag{13.26}$$

where k_i is the indicator for the original treatment assignment, not for the treatment after switching.

Model (13.24) or (13.25), however, does not take into account for the fact that treatment switch is typically based on prognosis and/or investigator's judgment. For example, a patient in one group may switch to another because he/she does not respond to the original assigned treatment. This may result in a somewhat optimal treatment assignment for the patient and a longer survival time than those patients who did not switch. Ignoring such a switching effect will lead to a biased assessment of the treatment effect. Shao, Chang, and Chow (2005) consider the following model conditional on S_i:

$$\tilde{T}_i \overset{d}{=} S_i + e^{\beta(1-2k_i)} w_{k,\eta}(S_i) (T_i - S_i), \tag{13.27}$$

where η is an unknown parameter vector and $w_{k,\eta}(S)$ are known functions of the switching time S when η and k are given. Typically, $w_{k,\eta}(S)$ should be close to 1 when S is near 0, i.e. the switching effect is negligible if switching occurs too early. Note that

$$\lim_{S \downarrow 0} w_{k,\eta}(S) = 1.$$

An example of the weight is given by

$$w_{k,\eta}(S) = exp\left(\eta_{k,0} S + \eta_{k,1} S^2\right), \tag{13.28}$$

where $\eta_{k,l}$ are unknown parameters.

Note that although a similar IPE method can be applied under model (13.27), it is not recommended for the following reason. If initial estimates of model parameters are obtained by solving the likelihood equation given in (13.28), then iteration does not increase the efficiency of estimates and hence adds unnecessary complexity for computation. On the other hand, if initial estimates are not solutions of the likelihood equation given in (13.28), then they are typically not efficient and the estimates obtained by IPE (if they converge) may not be as efficient as the solutions of the

likelihood equation (13.28). Thus, directly solving the likelihood equation (13.28) produces estimates that are either more efficient or computationally simpler than the IPE estimates. See the paper by Chao, Chang, and Chow (2005) for details.

13.5 Summary and discussions

We have introduced three different methods for modeling survival distribution with response-adaptive treatment switching: the mixed exponential model, the mixture of Wiener process, and the latent event model. The mixture of Wiener processes is very flexible and can model the covariates too, while the mixed exponential method is very simple and can be further developed to include baseline variates.

From an analysis point of view, the trial should be designed to allow treatment switching, but not crossover. Treatment crossover implies that patients in the control group will be allowed to switch to the test drug. In this case, the treatment difference is difficult to define. If a trial is designed to allow treatment switching but not crossover, then the comparison of the two groups (based on initial randomization) is easy to interpret. Suppose that if progressive disease (PD) is observed for a patient (from either group), the patient will switch to the best alternative treatment available on the market. This way, the control group represents the situation without the test drug, and the test group represents the situation with the test drug. The difference between these two is the patient's net health improvement by adding the test drug to the market. Of course, an oncology trial that does not allow for treatment crossover may be challenging with regard to patient enrollment and may also have some ethical issues.

Problem

13.1 Implement the three models discussed in this chapter using SAS or R and conduct a simulation study of the methods.

Chapter 14

Response-Adaptive Allocation Design

In this chapter, we will discuss the response-adaptive randomization/allocation designs, including the play-the-winner model and randomized play-the-winner model for two-arm trials, and the generalized urn model for multiple-arm trials with various endpoints. We will explore the properties of these adaptive designs and illustrate the methods with trial examples. Scholars including Zelen (1969), Wei and Durham (1978), Wei, Smythe and Lin (1990), Stallard, and Rosenberger (2002) and many others have contributed in this area.

14.1 Opportunities

Response-adaptive randomization or allocation is a randomization technique in which the allocation of patients to treatment groups is based on the response (outcome) of the previous patients. The purpose is to provide a better chance of randomizing the patients to a superior treatment group based on the knowledge about the treatment effect at the time of randomization. As a result, response-adaptive randomization takes ethical concerns into consideration. The well-known response-adaptive models include the play-the-winner (PW) model and the randomized play-the-winner (RPW) model.

14.1.1 *Play-the-Winner Model*

The play-the-winner (PW) model can be easily applied to clinical trials comparing two treatments (e.g., treatment A and treatment B) with binary outcomes (i.e., success or failure). For the PW model, it is assumed that the previous subject's outcome will be available before the next patient is randomized. The treatment assignment is based on the treatment response of the previous patient. If a patient responds to treatment A, then the next

patient will be assigned to treatment A. Similarly, if a patient responds to treatment B, then the next patient will be assigned to treatment B. If the assessment of the previous patients is not available, the treatment assignment can be based on the response assessment of the last available patient. It is obvious that this model lacks randomness.

14.1.2 *Randomized Play-the-Winner Model*

The randomized play-the-winner (RPW) model is a simple probabilistic model used to sequentially randomize subjects in a clinical trial (Wei and Durham, 1978; Coad and Rosenberger, 1999). The RPW model is useful for clinical trials comparing two treatments with binary outcomes. In the RPW model, it is assumed that the previous subject's outcome will be available before the next patient is randomized. At the start of the clinical trial, an urn contains a_0 balls representing treatment A and b_0 balls representing treatment B, where a_0 and b_0 are positive integers. We denote these balls as either type A or type B balls. When a subject is recruited, a ball is drawn and replaced. If it is a type A ball, the subject receives treatment A; if it is a type B ball, the subject receives treatment B. When a subject's outcome is available, the urn is updated as follows: A success on treatment A (B) or a failure on treatment B (A) will generate an additional a_1 (b_1) type-B balls in the urn. In this way, the urn builds up more balls representing the more successful treatment (Figure 14.1).

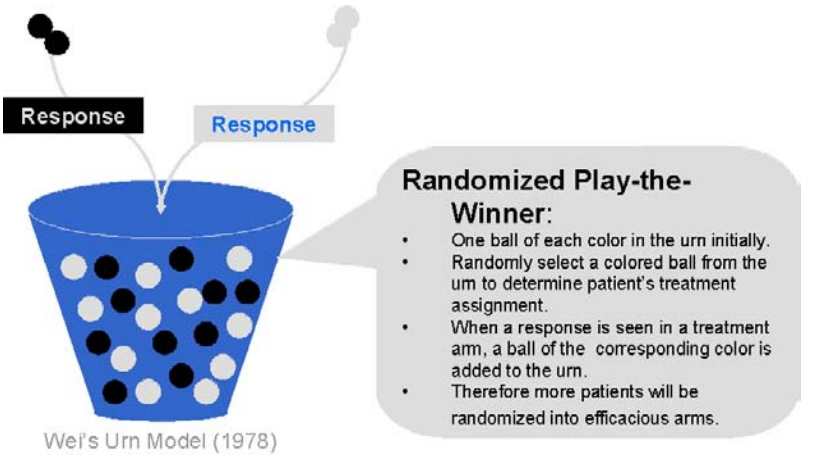


Figure 14.1: Randomized Play-the-Winner

There are some interesting asymptotic properties with RPW. Let N_a/N be the proportion of subjects assigned to treatment A out of N subjects.

Also, let $q_a = 1 - p_a$ and $q_b = 1 - p_b$ be the failure probabilities. Further, let F be the total number of failures. Then, we have (Wei and Durham, 1978)

$$\begin{cases} \lim\limits_{N\to\infty} \frac{N_a}{N_b} = \frac{q_b}{q_a}, \\ \lim\limits_{N\to\infty} \frac{N_a}{N} = \frac{q_b}{q_a+q_b}, \\ \lim\limits_{N\to\infty} \frac{F}{N} = \frac{2q_a q_b}{q_a+q_b}. \end{cases} \tag{14.1}$$

Since treatment assignment is based on response of the previous patient in RPW model, it is not optimized with respect to any clinical endpoint. It is desirable to randomize treatment assignment based on some optimal criteria such as minimizing the expected numbers treatment failures. This leads to the so-called optimal RPW model.

14.1.3 *Optimal RPW Model*

The optimal randomized play-winner model (ORPW) is intent to minimize the number of failures in the trial. There are three commonly used efficacy endpoints in clinic trials, namely, simple proportion difference ($p_a - p_b$), the relative risk (p_a/p_b) and the odds ratio ($p_a q_b/p_b q_a$), where $q_a = 1 - p_a$ and $q_b = 1 - p_b$ are failure rates. These can be estimated consistently by replacing p_a by $\hat{p}_a$ and p_b by $\hat{p}_b$, where $\hat{p}_a$ and $\hat{p}_b$ are the proportions of observed successes in treatment groups A and B, respectively. Suppose that we wish to find the optimal allocation $r = n_a/n_b$ such that it minimizes the expected number of treatment failures $n_a q_a + n_b q_b$ which is mathematically given by (Rosenberger and Lachin, 2002):

$$\begin{aligned} r^* &= \arg\min_r \{n_a q_a + n_b q_b\} \\ &= \arg\ \min_r \{\frac{r}{1+r} n\, q_a + \frac{1}{1+r} n\, q_b\}. \end{aligned} \tag{14.2}$$

For simple proportion difference, the asymptotic variance is given by

$$\frac{p_a q_a}{n_a} + \frac{p_b q_b}{n_b} = \frac{(1+r)(p_a\, q_a + r\, p_b\, q_b)}{n\, r} = K, \tag{14.3}$$

where K is some constant. Solving (14.3) for n yields

$$n = \frac{(1+r)(p_a\, q_a + r\, p_b\, q_b)}{rK}. \tag{14.4}$$

Substituting (14.4) into (14.2), we obtain:

$$r^* = \arg \min_r \left\{ \frac{(r\, p_a + q_b)(p_a q_a + r\, p_b q_b)}{r\, K} \right\}. \tag{14.5}$$

Taking the derivative of (14.5) with respect to r and equating to zero, we have

$$r^* = \left(\frac{p_a}{p_b} \right)^{\frac{1}{2}}.$$

Note that r^* does not depend on K.

Table 14.1: Asymptotic Variance with RPW

Measure	r^*	Asymptotic variance
Proportion difference	$\left(\frac{p_a}{p_b}\right)^{\frac{1}{2}}$	$\frac{p_a\, q_a}{n_a} + \frac{p_b\, q_b}{n_b}$
Relative risk	$\left(\frac{p_a}{p_b}\right)^{\frac{1}{2}} \left(\frac{q_b}{q_a}\right)$	$\frac{p_a\, q_b^2}{n_a q_a^3} + \frac{p_b\, q_b}{n_b q_a^2}$
Odds ratio	$\left(\frac{p_b}{p_a}\right)^{\frac{1}{2}} \left(\frac{q_b}{q_a}\right)$	$\frac{p_a\, q_b^2}{n_a q_a^3\, p_b^2} + \frac{p_b q_b}{n_b q_a^2 p_b^2}$

Source: Chow and Chang (2006, p.61).

Note that the limiting allocation for the RPW rule ($\frac{q_b}{q_a}$) is not optimal for any of the three measures and none of the optimal allocation rule yields Neyman allocation given by (Melfi and Page, 1998)

$$r^* = \left(\frac{p_a}{p_b} \frac{q_a}{q_b} \right)^{\frac{1}{2}}, \tag{14.6}$$

which minimizes the variance of the difference in sample proportions (Table 14.1). Note that Neyman allocation would be unethical when $p_a > p_b$ (i.e., more patients receive the inferior treatment).

Because the optimal allocation depends on the unknown binomial parameters, practically the unknown success probabilities in the optimal allocation rule can be replaced by the current estimate of the proportion of successes (i.e., $\hat{p}_{a,n}$ and $\hat{p}_{b,n}$) observed in each treatment group thus far.

14.2 Adaptive Design with RPW

We will use SAS Macro 14.1 to study the adaptive design using the randomized play-the-winner. There are different ways to design trial using RPW. We have seen that RPW can reduce the number of failures and increase the number of responses within a trial with fixed sample-size. However, it may

inflate α using a classic test statistic in conjunction with an unadjusted rejection region. In other words, it may reduce the power and sample-size has to be increased to retain the power. As a result, the increase in sample-size may lead to an increase of number of failures. Here are some immediate questions before we design a RPW trial. (1) How many analyses should be used? Should a full or group sequential design be used? (2) How to determine the four parameters in the randomization urn RPW(a_0, b_0, a_1, b_1)? (3) What test statistic should be used? (4) How to control α and calculate the power? (5) How to estimate the response rate in each group? We will use the SAS Macro 14.1, RPW, for two arm trial with binary endpoint to facilitate the discussion.

In SAS Macro 14.1, the initial numbers of balls in the urn are denoted by **a0** and **b0**. Next **a1** or **b1** balls added to the urn if a response is observed in arm A or arm B. The SAS variables are defined as follows: **RR1**, **RR2** = the response rates in group 1 and 2, respectively, **nSbjs** = total number of subjects (two groups combined), **nMin** (>0) = the minimum sample-size per group required to avoid an extreme imbalance situation, **nAnlys** = number of analyses (approximately an equal information-time design). All interim analyses are designed for randomization adjustment and only the final analysis for hypothesis testing. **aveP1** and **aveP2** = the average response rates in group 1 and 2, respectively. **Power** = probability of the test statistic > **Zc**. Note: **Zc** = function of (**nSbjs**, **nAnlys**, **a0**, **b0**, **a1**, **b1**, **nMin**).

>>SAS Macro 14.1: Randomized Play-the-Winner Design>>

```
%Macro RPW(nSims=100000, Zc=1.96, nSbjs=200, nAnlys=3,
     RR1=0.2, RR2=0.3, a0=1, b0=1, a1=1, b1=1, nMin=1);
Data RPW; Keep nSbjs aveP1 aveP2 Zc Power;
seed1=364; seed2=894; Power=0; aveP1=0; aveP2=0;
Do isim=1 to &nSims;
 nResp1=0; nResp2=0; a0=&a0; b0=&b0; Zc=&Zc; N1=0; N2=0;
 nSbjs=&nSbjs; nAnlys=&nAnlys; nMax=nSbjs-&nMin;
 a=a0; b=b0; r0=a/(a+b);
 Do iSbj=1 To nSbjs;
      If Mod(iSbj,Round(nSbjs/nAnlys))=0 Then r0=a/(a+b);
     rnAss=Ranuni(seed1);
     If (rnAss < r0 And N1<nMax) Or N2>=nMax Then
          Do;
               N1=N1+1; rnRep=Ranuni(seed2);
               if rnRep <=&RR1 Then Do;
                    nResp1=nResp1+1; a=a+&a1;
```

```
                End;
            End;
        Else
            Do;
                N2=N2+1; rnRep=Ranuni(seed2);
                If rnRep <=&RR2 Then Do;
                    nResp2=nResp2+1; b=b+&b1;
                End;
            End;
    End;
    p1=nResp1/N1; p2=nResp2/N2;
    aveP1=aveP1+p1/&nSims; aveP2=aveP2+p2/&nSims;
    sigma1=sqrt(p1*(1-p1)); sigma2=sqrt(p2*(1-p2));
    Sumscf=sigma1**2/(N1/(N1+N2))+sigma2**2/(N2/(N1+N2));
    TS = (p2-p1)*Sqrt((N1+N2)/sumscf);
    If TS>Zc Then Power=Power+1/&nSims;
End;
Output;
Run;
Proc Print data=RPW; Run;
%Mend RPW;
```

‹‹SAS‹‹

Example 14.1 Randomized Played-the-Winner Design

Suppose we are designing an oncology clinical study with tumor response as the primary endpoint. The response rate is estimated to be 0.3 in the control group and 0.5 in the test group. The response rate is 0.4 in both groups under the null condition. We want to design the trial with about 80% power at a one-sided α of 0.025.

We first check the type-I error of a classic two-group design with n = 200 (100/group) using the following SAS macro calls.

››SAS››

```
%RPW(Zc=1.96, nSbjs=200, nAnlys=200, RR1=0.4, RR2=0.4,
a0=1, b0=1,a1=0, b1=0, nMin=1);
```

‹‹SAS‹‹

Note that a1 = b1 = 0 represents the classic design. To calculate the power, we run the following code.

››SAS››

```
%RPW(Zc=1.96, nSbjs=200, nAnlys=200, RR1=0.3, RR2=0.5,
a0=1, b0=1,a1=0, b1=0, nMin=1);
‹‹SAS‹‹
```

The simulations indicate the power (or type-I error) is 83%. To determine the reject region or the critical point Z_c for a design with RPW$(1,1,1,1)$, we first use $Z_c = 1.96$, the critical point for the classic design in the following SAS macro call:

```
››SAS››
%RPW(Zc=1.96, nSbjs=200, nAnlys=200, RR1=0.4, RR2=0.4,
a0=1, b0=1,a1=1, b1=1, nMin=1);
‹‹SAS‹‹
```

The simulations indicate that the one-sided $\alpha = 0.055$, which is much larger than the target level 0.025. Therefore, using the SAS Macro 14.1 and trial-error method, we found that $Z_c = 2.7$ will give the power or $\alpha = 0.025$:

```
››SAS››
%RPW(Zc=2.7, nSbjs=200, nAnlys=200, RR1=0.4, RR2=0.4,
a0=1, b0=1,a1=1, b1=1, nMin=1);
‹‹SAS‹‹
```

Note the previous results are based on full sequential design, i.e., randomization is modified when each response assessment becomes available (assume no delayed response). Practically it is much easier to carry out a group sequential trial. For example, the trial can have five analyses and the randomization is modified at each of the four interim analyses. The final analysis is used for testing the null hypothesis of no treatment effect. We use the following SAS statement to find out that Z_c should be 2.05:

```
››SAS››
%RPW(Zc=2.05, nSbjs=200, nAnlys=5, RR1=0.4, RR2=0.4,
a0=1, b0=1,a1=1, b1=1, nMin=1);
‹‹SAS‹‹
```

Using the following SAS statement, we obtained the power that is 79%, 4% less than the classic design with the same sample-size:

```
››SAS››
%RPW(Zc=2.05, nSbjs=200, nAnlys=5, RR1=0.3, RR2=0.5,
a0=1, b0=1,a1=1, b1=1, nMin=1);
```

‹‹SAS‹‹

The results from the above eight different scenarios are summarized in Table 14.2.

Table 14.2: Simulation Results from RPW

Scenario	nSbjs	aveP1	aveP2	Zc	Power
1	200	0.400	0.400	1.96	0.0258
2	200	0.300	0.500	1.96	0.8312
3	200	0.381	0.381	1.96	0.0553
4	200	0.381	0.381	2.70	0.0256
5	200	0.395	0.396	2.05	0.0252
6	200	0.292	0.498	2.05	0.7908
7	200	0.396	0.396	2.035	0.0257
8	200	0.294	0.498	2.035	0.7983

Note: 100,000 Simulation runs.

Similarly we can study the characteristics of different urns by, for example, setting the parameters: a0 = 2, b0 = 2, a1 = 1, b1 = 1 for RPW (2,2,1,1). I left this for readers to practice.

14.3 General Response-Adaptive Randomization (RAR)

For other types of endpoints, we suggest the following allocation probability model:

$$\Pr\left(trt = i\right) = f\left(\hat{\mathbf{u}}\right), \tag{14.7}$$

where $\Pr\left(trt = i\right)$ is the probability of allocating the patient to the i^{th} group and the observed response vector $\hat{\mathbf{u}} = \{u_1, ..., u_M\}$.

We further suggest a specific function for f, i.e.,

$$\Pr(trt = i) \propto a_{0i} + b\,\hat{u}_i^m, \tag{14.8}$$

where $\hat{u}_i$ = the observed proportion, mean, number of events, or categorical score, and a_{0i} and b are constants.

14.3.1 *SAS Macro for M-Arm RAR with Binary Endpoint*

The response-adaptive randomization algorithm (14.8) has been implemented for the binary response in SAS Macro 14.2. The definitions of the

SAS variables are defined as follows: **nPts** = the total number of patients, **AveN{i}** = the average number of patients in the i^{th} arm, **AveU{i}** = the average response in the i^{th} arm, **PowerMax** = the power for testing the response difference between the arm with maximum response and the first arm, **nSims** = number of simulation runs, **nArms** = number of treatment arms, **Zc** = critical point for rejecting the null hypothesis. **a0{i}**, **b** and **m** are the parameters in the model (14.8).

››SAS Macro 14.2: Binary Response-Adaptive Randomization››

```
%Macro RARBin(nSims=1000, nPts=200, nArms=5, b=1, m=1, Zc=1.96);
 Data RARBin; Set DataIn;
 Keep nPts AveN1-AveN&nArms AveU1-AveU&nArms PowerMax;
 Array Ns{&nArms}; Array uObs{&nArms}; Array rP{&nArms};
 Array nRsps{&nArms}; Array a0{&nArms}; Array CrP{&nArms};
 Array us{&nArms}; Array AveU{&nArms}; Array AveN{&nArms};
 PowerMax=0; nArms=&nArms; nPts=&nPts;
 Do i=1 To nArms; AveU{i}=0; AveN{i}=0; End;
 Do isim=1 to &nSims;
    Do i=1 To nArms; nRsps{i}=0; uObs{i}=0; Ns{i}=0; Crp{i}=0; End;
    Do iSubject=1 to nPts;
      Do i=1 To nArms; rP{i}=a0{i}+&b*uObs{i}**&m; End;
      Suma=0; Do i=1 To nArms; Suma=Suma+rP{i}; End;
      Do i=1 To nArms; rP{i}=rP{i}/Suma; End;
      Do iArm=1 To nArms; CrP{iArm}=0;
           Do i=1 To iArm; CrP{iArm}=CrP{iArm}+rP{i}; End;
      End;
      rn=ranuni(5236); cArm=1;
      Do iArm=2 To nArms;
           IF CrP{iArm-1}<rn<CrP{iArm} Then cArm=iArm;
      End;
      Ns(cArm)= Ns(cArm)+1;
      * For Binary response;
      If ranuni(8364)<us{cArm} Then nRsps{cArm}=nRsps{cArm}+1;
      Do i=1 To nArms; uObs{i}=nRsps{i}/max(Ns{i},1); End;
    End;
  uMax=uObs{1};
    Do i=1 to &nArms; If uObs{i}>=uMax Then iMax=i; End;
    Se2=0;
    Do i=1 to &nArms;
     Se2=Se2+uObs{i}*(1-uObs{i})/max(Ns{i},1)*2/nArms;
```

```
    End;
    TSmax=(uObs{iMax}-uObs{1})*(Ns(1)+Ns{iMax})/2
                              /(nPts/nArms)/Se2**0.5;
    If TSmax>&Zc then PowerMax=PowerMax+1/&nSims;
    Do i=1 To nArms;
      AveU{i}=AveU{i}+uObs{i}/&nSims;
      AveN{i}=AveN{i}+Ns{i}/&nSims;
    End;
  End;
  Output;
  Run;
  Proc Print Data=RARBin; Run;
  %Mend RARBin;
```

‹‹**SAS**‹‹

Examples of RAR designs using SAS Macro 14.2 are presented as follows:

```
>>SAS>>
Title "Checking Alpha for 2-Group with Classic Design";
Data DataIn;
Array a0{2} (1,1); Array us{2} (0.2,0.2);
%RARBin(nSims=10000, nPts=160, nArms=2, b=0, m=1, Zc=1.96); Run;

Title "Checking Alpha for 2-Group RAR Design";
Data DataIn;
Array a0{2} (1,1); Array us{2} (0.2,0.2);
%RARBin(nSims=100000, nPts=160, nArms=2, b=1, m=1, Zc=2.0); Run;

Title "Power with 2-Group RAR Design";
Data DataIn;
Array a0{2} (1,1); Array us{2} (0.2,0.4);
%RARBin(nSims=100000, nPts=160, nArms=2, b=1, m=1, Zc=2.0); Run;

Title "Checking Alpha for 3-Group RAR Design";
Data DataIn;
Array a0{3} (1,1,1); Array us{3} (0.2,0.2,0.2);
%RARBin(nSims=100000, nPts=160, nArms=3, b=1, m=1, Zc=2.08); Run;

Title "Power with 3-Group RAR Design";
Data DataIn;
Array a0{3} (1,1,1); Array us{3} (0.2,0.3,0.5);
```

```
%RARBin(nSims=100000, nPts=160, nArms=3, b=1, m=1, Zc=2.08); Run;
```
‹‹**SAS**‹‹

14.3.2 *SAS Macro for M-Arm RAR with Normal Endpoint*

Algorithm (14.8) has also been implemented for normal response in SAS Macro 14.3.

››SAS Macro 14.3: Normal Response-Adaptive Randomization››

```
%Macro RARNor(nSims=100000, nPts=100, nArms=5, b=1, m=1,
                                        CrtMax=1.96);
Data RARNor; Set DataIn;
Keep nPts AveN1-AveN&nArms AveU1-AveU&nArms PowerMax;
Array Ns{&nArms}; Array AveN{&nArms}; Array uObs{&nArms};
Array rP{&nArms}; Array AveU{&nArms}; Array cuObs{&nArms};
Array a0{&nArms}; Array CrP{&nArms};
Array us{&nArms}; Array s{&nArms};
PowerMax=0; nArms=&nArms; nPts=&nPts;
Do i=1 To nArms; AveU{i}=0; AveN{i}=0; End;
Do isim=1 to &nSims;
 Do i=1 To nArms; cuObs{i}=0; uObs{i}=0; Ns{i}=0; Crp{i}=0; End;
 Do iSubject=1 to nPts;
      Do i=1 To nArms; rP{i}=a0{i}+&b*uObs{i}**&m; End;
      Suma=0; Do i=1 To nArms; Suma=Suma+rP{i}; End;
      Do i=1 To nArms; rP{i}=rP{i}/Suma; End;
     Do iArm=1 To nArms; CrP{iArm}=0;
          Do i=1 To iArm; CrP{iArm}=CrP{iArm}+rP{i}; End;
     End;
     rn=ranuni(5361); cArm=1;
 Do iArm=2 To nArms;
          IF CrP{iArm-1}<rn<CrP{iArm} Then cArm=iArm;
 End;
     Ns(cArm)= Ns(cArm)+1;
 * For Normal response;
     u=Rannor(361)*s{cArm}+us{cArm};
 cuObs{cArm}=cuObs{cArm}+u;
     Do i=1 To nArms; uObs{i}=cuObs{i}/max(Ns{i},1); End;
 End;
 se2=0;
```

```
* Assume sigma unknown for simplicity;
Do i=1 To nArms; se2=se2+s{i}**2/max(Ns{i},1)*2/nArms; End;
uMax=uObs{1};
Do i=1 To &nArms; If uObs{i}>=uMax Then iMax=i; End;
TSmax=(uObs{iMax}-uObs{1})*(Ns(1)+Ns{iMax})/2
                          /(nPts/nArms)/se2**0.5;
If TSmax>&CrtMax then PowerMax=PowerMax+1/&nSims;
Do i=1 To nArms;
    AveU{i}=AveU{i}+uObs{i}/&nSims;
    AveN{i}=AveN{i}+Ns{i}/&nSims;
End;
End;
Output;
Run;
Proc Print data=RARNor; Run;
%Mend RARNor;
```
‹‹**SAS**‹‹

Example 14.2 Adaptive Randomization with Normal Endpoint

The objective of this trial in asthma patients is to confirm sustained treatment effect, measured as FEV1 change from baseline to the 1-year of treatment. Initially, patients are equally randomized to four doses of the new compound and a placebo. Based on early studies, the estimated FEV1 change at week 4 are 6%, 12%, 13%, 14%, and 15% (with pooled standard deviation 18%) for the placebo, dose level 1, 2, 3, and 4, respectively.

Using the following SAS macro calls, we can determine that the rejection region is $(0.848, +\infty)$. The power is 84% with a total of 375 subjects and 73% with 285 subjects, while the power is 90% for the drop-loser design with an expected sample-size of 285 in Example 11.1 (Table 11.3).

```
››SAS››
Data DataIn;
Array a0{5} (1, 1, 1, 1, 1); Array us{5} (0.06, 0.06, 0.06, 0.06, 0.06);
Array s{5} (.18, .18, .18, .18, .18);
%RARNor(nPts=375, nArms=5, b=1, m=1, CrtMax=2.01);

Data DataIn;
Array a0{5} (1, 1, 1, 1, 1); Array us{5} (0.06, 0.12, 0.13, 0.14, 0.15);
Array s{5} (.18, .18, .18, .18, .18);
%RARNor(nPts=375, nArms=5, b=1, m=1, CrtMax=2.01);
```
‹‹**SAS**‹‹

```
>>SAS>>
Data DataIn;
 Array a0{5} (1, 1, 1, 1, 1); Array us{5} (0.06, 0.06, 0.06, 0.06, 0.06);
 Array s{5} (.18, .18, .18, .18, .18);
 %RARNor(nPts=285, nArms=5, b=1, m=1, CrtMax=1.995);
 Data DataIn;
 Array a0{5} (1, 1, 1, 1, 1); Array us{5} (0.06, 0.12, 0.13, 0.14, 0.15);
 Array s{5} (.18, .18, .18, .18, .18);
 %RARNor(nPts=285, nArms=5, b=1, m=1, CrtMax=1.995);
<<SAS<<
```

14.3.3 *RAR for General Adaptive Designs*

Many adaptive designs can be viewed as response-adaptive randomization design. We are going to illustrate this with the classic group sequential design and the drop-loser design.

For two-arm group sequential design, the treatment allocation probability to the i^{th} arm at the k^{th} stage is given by

$$\Pr(trt = i; k) = \frac{H(p_k - \alpha_k) + H(\beta_k - p_k)}{2} - \frac{1}{2}, \tag{14.9}$$

where $i = 1, 2$, α_k and β_k are the stopping boundaries at the k^{th} stage and the step-function is defined as

$$H(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}. \tag{14.10}$$

For drop-loser designs, if the dropping-criterion is based on the maximum difference among the observed treatment $\hat{u}_{\max} - \hat{u}_{\min}$ at the interim analysis, then the treatment allocation probability to the i^{th} arm at interim analysis is given by

$$\Pr(trt = i) = \frac{H(\hat{u}_i - \hat{u}_{\max} + \delta_{NI})}{\sum_{i=1}^{M} H(\hat{u}_i - \hat{u}_{\max} + \delta_{NI})}, \tag{14.11}$$

where $\hat{u}_i$ is the observed response in the i^{th} arm, and δ_{NI} is the noninferiority margin. $M =$ the total number of arms in the study.

14.4 Summary and Discussion

Response-adaptive randomization (RAR) was initially proposed to reduce the number of failures in a trial; however, the overall gain is limited because (1) power is lost as compared to the uniformly-most powerful design, and (2) the reduction in number of failures can diminish due to significantly delayed responses. RAR may (although it is unlikely) delay patient enrollment because of the fact that patients enrolled later will have a better chance of being assigned to a better treatment group. If there is heterogeneity in patient enrollment over time (e.g., sicker patients tend to enroll earlier because they cannot wait for long), a bias will be introduced.

RAR may be useful in phase II/III combination studies, where at the early stages, RAR is used to seamlessly select superior arms. In practice, group (rather than full) sequential response-randomization may be used, where the response data will be unblinded at several prescheduled times. In summary, RAR is practically more useful in drop-loser designs than two-group designs.

Problem

14.1 Modify Macro 14.3 so that it can calculate and output the mean response per person per trial.

Chapter 15

Adaptive Dose Finding Design

In this chapter, we will introduce two commonly used approaches for oncology dose-escalation trials: (1) the algorithm-based escalation rules, and (2) model-based approach. The second approach can be frequentist or Bayesian-based response-adaptive method and can be used in any dose-response trials.

15.1 Oncology Dose-Escalation Trial

For non-life-threatening diseases, since the expected drug toxicity is mild and can be controlled without harm, phase I trials are usually conducted on healthy or normal volunteers. In life-threatening diseases such as cancer and AIDS, phase I studies are conducted with a limited number of patients due to (1) the aggressiveness and possible harmfulness of cytotoxic treatments, (2) possible systemic treatment effects, and (3) the high interest in the new drug's efficacy in those patients directly.

Drug toxicity is considered as tolerable if the toxicity is manageable and reversible. The standardization of the level of drug toxicity is the Common Toxicity Criteria (CTC) developed by the United States National Cancer Institute (NCI). Any adverse event (AE) related to treatment of CTC category of Grade 3 and higher is often considered a dose-limiting toxicity (DLT). The maximum tolerated dose (MTD) is defined as the maximum dose with a DLT rate that is no more frequent than a predetermined value.

15.1.1 *Dose Level Selection*

The initial dose given to the first patients in a phase I study should be low enough to avoid severe toxicity. The commonly used starting dose is the dose at which 10% mortality (LD_{10}) occurs in mice. The subsequent dose levels are usually selected based on the following multiplicative

set, $x_i = f_{i-1}x_{i-1}$ $(i = 1, 2, ...k)$, where f_i is called the dose escalation factor. The highest dose level should be selected such that it covers the biologically active dose, but remains lower than the toxic dose. A popular dose-escalation factor is called the Fibonacci sequence (2, 1.5, 1.67, 1.60, 1.63, 1.62, 1.62, ...) and modified Fibonacci sequence (2, 1.65, 1.52, 1.40, 1.33, 1.33, ...). Note that the latter is monotonic sequence, hence more appropriate than the former.

15.1.2 *Traditional Escalation Rules*

There are usually 5 to 10 predetermined dose levels with modified Fibonacci sequence in a dose escalation study. Patients are treated with the lowest dose first and then gradually escalated to higher doses if there is no major safety concern. The rules for dose escalation are predetermined. The commonly employed set of dose escalation rules is the traditional escalation rules (TER), also known as the "3 + 3" rule. The "3 + 3" rule says to enter three patients at a new dose level and enter another 3 patients when only one DLT is observed. The assessment of the six patients will be performed to determine whether the trial should be stopped at that level or to increase the dose. Basically, there are two types of the "3 + 3" rules, namely, TER and strict TER (or STER). TER does not allow dose de-escalation, but STER does. The "3+3" TER and STER can be generalized to "A+B" TER and STER.

To introduce the $A+B$ escalation rule, let A, B, C, D, and E be integers. The notation A/B indicates that there are A toxicity incidences out of B subjects and $>A/B$ means that there are more than A toxicity incidences out of B subjects. We assume that there are K predefined doses and let p_i be the probability of observing a DLT at dose level i for $1 \leq i \leq K$.

A + B Escalation without Dose De-escalation:

The general $A + B$ designs without dose de-escalation can be described as follows. Suppose that there are A patients at dose level i. If fewer than C/A patients have DLTs, then the dose is escalated to the next dose level $i + 1$. If more than D/A (where $D \geq C$) patients have DLTs, then the previous dose $i - 1$ will be considered the MTD. If no fewer than C/A but no more than D/A patients have DLTs, B more patients are treated at this dose level i. If no more than E (where $E \geq D$) of the total of $A+B$ patients experience DLTs, then the dose is escalated. If more than E of the total of $A + B$ patients have DLT, then the previous dose $i - 1$ will be considered the MTD. It can be seen that the traditional "3 + 3" design without dose de-escalation is a special case of the general $A + B$ design with $A = B = 3$ and $C = D = E = 1$. The closed forms of operating characteristics (Lin

and Shih, 2001) are given below.

Under the general $A + B$ design without de-escalation, the probability of concluding that MTD has reached at dose i is given by

$$P_i^* = P(MTD = dose\ i) = P\begin{pmatrix}\text{escalation at dose } \leq i \text{ and} \\ \text{stop escalation at dose } i+1\end{pmatrix}$$
$$= (1 - P_0^{i+1} - Q_0^{i+1}) \left(\prod_{j=1}^{i} (P_0^j + Q_0^j) \right),\ 1 \leq i < K, \tag{15.1}$$

where

$$P_0^j = \sum_{k=0}^{C-1} \binom{A}{k} p_j^k (1 - p_j)^{A-k},$$

and

$$Q_0^j = \sum_{k=C}^{D} \sum_{m=0}^{E-k} \binom{A}{k} p_j^k (1 - p_j)^{A-k} \binom{B}{m} p_j^m (1 - p_j)^{B-m}.$$

Here p_j is the toxicity (DLT) rate at dose level j.

An overshoot is defined as an attempt to escalate to a dose level at the highest level planned, while an undershoot is defined as an attempt to de-escalate to a dose level at a lower dose than the starting dose level. Thus, the probability of undershoot is given by

$$P_1^* = P(MTD < dose\ 1) = (1 - P_0^1 - Q_0^1), \tag{15.2}$$

and the probability of overshoot is given by

$$P_n^* = P(MTD \geq dose\ K) = \Pi_{j=1}^{K} (P_0^j + Q_0^j). \tag{15.3}$$

The expected number of patients at dose level j is given by

$$N_j = \sum_{i=0}^{K-1} N_{ji} P_i^*, \tag{15.4}$$

where

$$N_{ji} = \begin{cases} \frac{AP_0^j + (A+B)Q_0^j}{P_0^j + Q_0^j} & \text{if } j < i+1 \\ \frac{A(1-P_0^j - P_1^j) + (A+B)(P_1^j - Q_0^j)}{1-P_0^j - Q_0^j} & \text{if } j = i+1 \\ 0 & \text{if } j > i+1 \end{cases}.$$

Note that without consideration of undershoots and overshoots, the expected number of DLTs at dose i can be obtained as $N_i p_i$. As a result, the total expected number of DLTs for the trial is given by $\sum_{i=1}^{K} N_i p_i$.

A + B Escalation with Dose De-escalation:

Basically, the general $A + B$ design with dose de-escalation is similar to the design without dose de-escalation. However, it permits more patients to be treated at a lower dose (i.e. dose de-escalation) when excessive DLT incidences occur at the current dose level. The dose de-escalation occurs when more than D/A (where $D \geq C$) or more than $E/(A + B)$ patients have DLTs at dose level i. In this case, B more patients will be treated at dose level $i-1$, provided that only A patients have been previously treated at this prior dose. If more than A patients have already been treated previously, then dose $i - 1$ is the MTD. The de-escalation may continue to the next dose level $i - 2$ and so on if necessary. The closed forms of operating characteristics are given by Lin and Shih (2001) as follows.

The probability of concluding that the MTD has been reached at dose i is given by

$$\begin{aligned} P_i^* &= P(MTD = dose\ i) = P\begin{pmatrix} \text{escalation at dose } \leq i \text{ and} \\ \text{stop escalation at dose } i+1 \end{pmatrix} \\ &= \sum_{k=i+1}^{K} p_{ik}, \end{aligned} \tag{15.5}$$

where

$$p_{ik} = (Q_0^i + Q_1^i)(1 - P_0^k - Q_0^k) \left(\prod_{j=1}^{i-1} (P_0^j + Q_0^j) \right) \prod_{j=i+1}^{k-1} Q_2^j, \tag{15.6}$$

$$Q_1^j = \sum_{k=0}^{C-1} \sum_{m=0}^{E-k} \binom{A}{k} p_j^k (1-p_j)^{A-k} \binom{B}{m} p_j^m (1-p_j)^{B-m}, \tag{15.7}$$

and

$$Q_2^j = \sum_{k=0}^{C-1} \sum_{m=E+1-k}^{B} \binom{A}{k} p_j^k (1-p_j)^{A-k} \binom{B}{m} p_j^m (1-p_j)^{B-m}, \quad (15.8)$$

Also, the probability of undershooting is given by

$$P_1^* = P(MTD < dose\ 1) = \sum_{k=1}^{K} \{ \left(\Pi_{j=1}^{k-1} Q_2^j \right) (1 - P_0^k - Q_0^k) \}, \quad (15.9)$$

and the probability of overshooting is

$$P_K^* = P(MTD \geq dose\ K) = \Pi_{j=1}^{K} (P_0^j + Q_0^j). \quad (15.10)$$

The expected number of patients at dose level j is given by

$$N_j = N_{jK} P_K^* + \sum_{i=0}^{K-1} \sum_{k=i+1}^{K} N_{jik} p_{ik}, \quad (15.11)$$

where

$$N_{jn} = \frac{A P_0^j + (A+B) Q_0^j}{P_0^j + Q_0^j}, \quad (15.12)$$

$$N_{jik} = \begin{cases} \frac{A P_0^j + (A+B) Q_0^j}{P_0^j + Q_0^j} & \text{if} \quad j < i \\ A + B & \text{if } i \leq j < k \\ \frac{A(1-P_0^j - P_1^j) + (A+B)(P_1^j - Q_0^j)}{1 - P_0^j - Q_0^j} & \text{if} \quad j = k \\ 0 & \text{if} \quad j > k \end{cases}, \quad (15.13)$$

and

$$P_1^j = \sum_{i=C}^{D} \binom{A}{k} p_j^k (1-p_j)^{A-k}. \quad (15.14)$$

Consequently, the total number of expected DLTs is given by $\sum_{i=1}^{K} N_i p_i$.

15.1.3 *Simulations Using SAS Macro*

The objective of SAS Macro 15.1 is to simulate the 3+3 traditional escalation. The SAS variables are defined as follows: **nSims** = number of simulation runs, **nLevels** = number of dose levels, **DeEs** = "true" means that it allows for dose de-escalation, otherwise, it does not. **AveMTD** =

average observed MTD, **AveNPts** = average number of patient per trial, **AveNRsps** = average number of responses (DLTs) in a trial.

››SAS Macro 15.1: 3+3 Dose-Escalation Design››

```
%Macro TER3p3(nSims=10000, DeEs="true", nLevels=10);
Data TER; Set dInput; Keep AveMTD SdMTD AveNPts AveNRsps;
Array nPts{&nLevels}; Array nRsps{&nLevels}; Array RspRates{&nLevels};
AveMTD=0; VarMTD=0; AveNPts=0; AveNRsps=0; nLevels=&nLevels;
Do iSim=1 to &nSims;
  Do i=1 To nLevels; nPts{i}=0; nRsps{i}=0; End;
  seedn=Round((Ranuni(281)*100000000));
  iLevel=1; TotPts=0; TotRsps=0;
Looper:
  If iLevel>&nLevels | iLevel<1 Then Goto Finisher;
  nPts{iLevel}=nPts{iLevel}+3;
  rspRate=RspRates{iLevel};
  Rsp=RANBIN(seedn,3,rspRate);
  nRsps{iLevel}=nRsps{iLevel}+Rsp;
  TotPts=TotPts+3; TotRsps=TotRsps+Rsp;
  If nPts(iLevel)=3 & nRsps{iLevel}=0 Then Do;
     iLevel=iLevel+1;
     Goto Looper;
  End;
  If nPts(iLevel)=3 & nRsps{iLevel}=1 Then Goto Looper;
  If nPts(iLevel)=3 & nRsps{iLevel}>1 Then Do;
     If &DeEs="false" | iLevel=1 Then Goto Finisher;
     iLevel=iLevel-1;
     Goto Looper;
  End;
  If nPts(iLevel)=6 & nRsps{iLevel}<1 Then Do;
     iLevel=iLevel+1;
     Goto Looper;
  End;
 Finisher:
  MTD=Min(iLevel, nLevels);
  AveMTD=AveMTD+MTD/&nSims;
  VarMTD=VarMTD+MTD**2/&nSims;
  AveNPts=AveNPts+totPts/&nSims;
  AveNRsps=AveNRsps+TotRsps/&nSims;
End;
  SdMTD=(VarMTD-AveMTD**2)**0.5;
```

```
Output;
Run;
Proc Print Data=TER; Run;
%Mend TER3p3;
```
‹‹SAS‹‹

We will show you later in Example 15.1 how to use this macro.

15.2 Continual Reassessment Method (CRM)

The continual reassessment method (CRM) is a model approach, in which the parameters in the model for the response are continually updated based on the observed response data. The method for updating the parameters can be either frequentist or Bayesian approach.

CRM was initially proposed by O'Quigley (O'Quigley, et al., 1990; O'Quigley and Shen, 1996; Babb and Rogatko, 2004) for oncology dose-escalation trial. However, it can be extended to other types of trials (Chang and Chow 2006). In CRM, the dose-response relationship is continually re-assessed based on accumulative data collected from the trial. The next patient who enters the trial is then assigned to the currently estimated MTD level. This approach is more efficient than TER with respect to finding the MTD.

Let's denote prior distribution by $\pi(\theta)$, and the sample distribution by $f(x|\theta)$. In Bayesian approach for CRM, there are four basic elements:

(1) the joint distribution of (θ, x) given by

$$\varphi(\theta, x) = f(x|\theta)\,\pi(\theta), \tag{15.15}$$

(2) the marginal distribution of x given by

$$m(x) = \int \varphi(\theta, x)\, d\theta = \int f(x|\theta)\,\pi(\theta)\, d\theta, \tag{15.16}$$

(3) the posterior distribution of θ given by Bayes' formula

$$\pi(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{m(x)}, \tag{15.17}$$

(4) the predictive probability distribution given by

$$P(y|x) = \int P(x|y, \theta)\,\pi(\theta|x)\, d\theta. \tag{15.18}$$

15.2.1 *Probability Model for Dose-Response*

Let x be the dose or dose level, and $p(x)$ be the probability of response or response rate. The commonly used model for dose-response is logistic model (Figure 15.1).

$$p(x) = [1 + b\exp(-ax)]^{-1}, \tag{15.19}$$

where b is usually a predetermined constant and a is a parameter to be updated based on observed data.

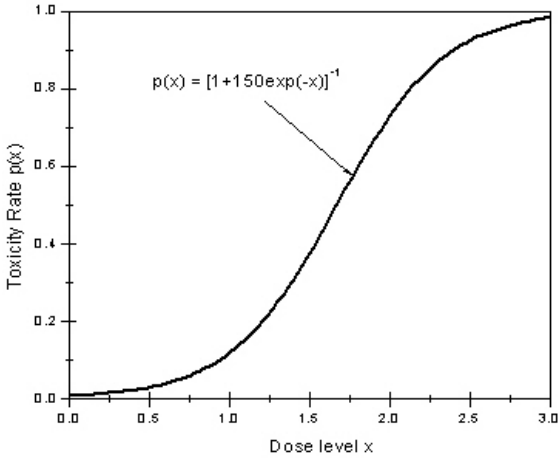


Figure 15.1: Logistic Toxicity Model

15.2.2 *Prior Distribution of Parameter*

The Bayesian approach requires the specification of prior probability distribution of the unknown parameter a.

$$a \sim g_0(a), \tag{15.20}$$

where $g_0(a)$ is the prior probability.

When there is very limited knowledge about the prior is available, non-informative prior can be used.

Likelihood Function

The next step is to construct the likelihood function. Given n observations with y_i $(i = 1, ..., n)$ associated with dose x_{m_i}, the likelihood function

can be written as

$$f_n(\mathbf{r}\,|a) = \prod_{i=1}^{n} \left[p(x_{m_i})\right]^{r_i} \left[1 - p(x_{m_i})\right]^{1-r_i}, \tag{15.21}$$

where

$$r_i = \begin{cases} 1, & \text{if response observed for } x_{m_i} \\ 0, & \text{otherwise} \end{cases}. \tag{15.22}$$

15.2.3 *Reassessment of Parameter*

The key is to estimate the parameter a in the response model (15.19). An initial assumption or a prior about the parameter is necessary in order to assign patients to the dose level based on the dose-toxicity relationship. This estimation of a is continually updated based on the cumulative response data observed from the trial thus far. The estimation method can be a Bayesian or frequentist approach. For Bayesian approach, it leads to the posterior distribution of a. For frequentist approaches, maximum likelihood estimate or least square estimate can be used.

Bayesian approach

For Bayesian approach, the posterior probability of parameter a can be obtained as follows

$$g_n(a|\mathbf{r}) = \frac{f_n(\mathbf{r}|a) g_0\,(a)}{\int f_n(\mathbf{r}|a) g_0(a)\, da} \tag{15.23}$$

or

$$g_n(a|\mathbf{r}) = \frac{\left[p_n\,(a)\right]^{r_n} \left[1 - p_n\,(a)\right]^{1-r_n} g_{n-1}\,(a)}{\int \left[p_n\,(a)\right]^{r_n} \left[1 - p_n\,(a)\right]^{1-r_n} g_{n-1}(a)\, da}, \tag{15.24}$$

where $p_n\,(a) = p(x_{m_n})$ is the response rate at the dose level, at which the n^{th} patient is treated.

After having obtained $g_n(a|\mathbf{r})$, we can update the predictive probability using

$$p(x) = \int \left[1 + b \exp(-ax)\right]^{-1} g_n(a|\mathbf{r})\, da. \tag{15.25}$$

Maximum likelihood approach

Note that Bayesian approach is computationally intensive. Alternatively, we may consider a frequentist approach to simplify the calculation. The maximum likelihood estimate of the parameters is given by

$$\hat{a} = \arg\max_{a} \{f_n(\mathbf{r} \mid a)\}. \tag{15.26}$$

Note that the MLE $\hat{a}$ is only available after both responders and non-responders are observed.

After having obtained $\hat{a}$, we can update the dose-response model or the predictive probability using

$$p(x) = [1 + b\exp(-\hat{a}x)]^{-1}. \tag{15.27}$$

15.2.4 *Assignment of Next Patient*

The updated dose-toxicity model is usually used to choose the dose level for the next patient. In other words, the next patient enrolled in the trial is assigned to the currently estimated MTD based on dose-response model (15.25) or (15.27). Practically, this assignment is subject to safety constraints such as limited dose jump. Assignment of patient to the most updated MTD is intuitive. This way, majority of the patients will be assigned to the dose levels near MTD, which allows for a more precise estimation of MTD with a minimal number of patients.

15.2.5 *Simulations of CRM*

SAS Macro 15.2, CRM, is developed for simulating the trial using CRM, where the logistic response model (5.19) is used. The input SAS variables are defined as follows: **nSims** = number of simulations, **nPts** = total number of patients, **nLevels** = number of dose levels, **b** = model parameter in (15.19), **aMin** and **aMax** = the upper and lower limits for prior on the parameter a, **MTRate** = the rate defined for MTD, **nIntPts** = number of intervals for numerical integration in calculating the posterior, **g{i}** = prior distribution of the model parameter a, **RRo{i}** = true response rates, and **doses{i}** = dose amount. These three arrays should be in the dataset naming **DInput**. The key output variables are: **AveMTD** = average MTD (simulated), **SdMTD** = standard deviation of MTDs (simulated).

>>SAS Macro 15.2: Continual Reassessment Method>>

```
%Macro CRM(nSims=100, nPts=30, nLevels=10, b=100,
    aMin=0.1, aMax=0.3, MTRate=0.3, nIntPts=100);
Data CRM; Set DInput; Keep nPts nLevels AveMTD SdMTD DLTS;
Array nPtsAt{&nLevels}; Array nRsps{&nLevels}; Array g{&nIntPts};
Array Doses{&nLevels}; Array RRo{&nLevels}; Array RR{&nLevels};
```

```
seed=2736; nLevels=&nLevels; nPts=&nPts; DLTs=0;
AveMTD=0;VarMTD=0; dx=(&aMax-&aMin)/&nIntPts;
Do iSim=1 to &nSims;
 Do i=1 To nLevels; nPtsAt{i}=0; nRsps{i}=0; End;
 iLevel=1;
 Do iPtient=1 To nPts;
 iLevel=min(iLevel, &nLevels); Rate=RRo{iLevel};
 nPtsAt{iLevel}=nPtsAt{iLevel}+1;
     r=Ranbin(seed,1,Rate); nRsps{iLevel}=nRsps{iLevel}+r;
** Posterior distribution of a;
     c=0;
     Do k=1 To &nIntPts;
          ak=&aMin+k*dx;
          Rate=1/(1+&b*Exp(-ak*doses{iLevel}));
          If r>0 Then L=Rate; Else L=(1-Rate);
          g{k}=L*g{k}; c=c+g{k}*dx;
     End;
     Do k=1 to &nIntPts; g{k}=g{k}/c; End;
** Predict response rate and current MTD;
      MTD=iLevel; MinDR=1;
      Do i=1 To nLevels;
           RR{i}=0;
          Do k=1 To &nIntPts;
          ak=&aMin+k*dx;
      RR{i}= RR{i}+1/(1+&b*Exp(-ak*doses{i}))*g{k}*dx;
          End;
          DR=Abs(&MTRate-RR{i});
          If .<DR <MinDR Then
               Do; MinDR = DR; iLevel=i; MTD=i; End;
     End;
 End;
 Do i=1 To nLevels;
     DLTs=DLTs+nRsps{i}/&nSims;
 End;
 AveMTD=AveMTD+MTD/&nSims;
 VarMTD=VarMTD+MTD**2/&nSims;
End;
SdMTD=(VarMTD-AveMTD**2)**0.5;
Output;
Run;
Proc Print Data=CRM; run;
```

```
%Mend CRM;
‹‹SAS‹‹
```

15.2.6 *Evaluation of Dose-Escalation Design*

There are advantages and disadvantages with different dose escalation schemes. For example, the traditional 3+3 escalation is easy to apply but the MTD estimation is usually biased, especially when there are many dose levels. The criteria for evaluation of escalation schemes are listed as follows: the number of DLTs, number of patients, number of patients dosed above MTD, accuracy, and precision.

Before a phase I trial is initiated, the following design characteristics should be checked and defined in the study protocol: (1) starting dose, (2) dose levels, (3) prior information on the MTD, (4) toxicity model, (5) escalation rule, (6) stopping rule, and (7) rules for completion of the sample-size when stopping.

Example 15.1 Adaptive Dose-Finding for Prostate Cancer Trial

A trial is designed to establish the dose-toxicity relationship and identify MTD for a compound in patients with metastatic androgen independent prostate cancer. Based on preclinical data, the estimated MTD is 230 mg/m^2. The modified Fibonacci sequence is chosen for the dose levels (in Table 15.1). There are 8 dose levels anticipated, but more dose levels can be added if necessary. The initial dose level is 30 mg/m^2, which is 1/10 of the minimal effective dose level (mg/m^2) for 10% deaths (MELD10) of the mouse after verification that no lethal and no life-threatening effects were seen in another species. The toxicity rate (DLT rate) at MTD is defined for this indication as 17%.

Table 15.1: Dose Levels and DLT Rates

Dose level i	1	2	3	4	5	6	7	8
Dose x	30	45	68	101	152	228	342	513
DLT rate	0.01	0.02	0.03	0.05	0.12	0.17	0.22	0.4

The SAS macro calls for TER and STER designs are presented as follows:

```
››SAS››
Title "3 + 3 TER and SER Designs";
Data dInput;
```

```
Array RspRates{8}(0.01,0.02,0.03,0.05,0.12,0.17,0.22,0.4);;
%TER3p3(nSims=100000, DeEs="true", nLevels=8);
%TER3p3(nSims=100000, DeEs="false", nLevels=8);
 run;
```
‹‹SAS‹‹

The SAS macro calls for CRM design are presented as follows:

››SAS››
```
Title "Bayesian CRM Design";
Data DInput;
Array g{100}; Array RRo{8}(.01, .02, .03, .05, .12, .17, .22, .4);
Array Doses{8} (30, 45, 67, 100, 150, 230, 340, 510);
 Do k=1 To 100; g{k}=1; End; * Flat prior;
 Do i=1 To 8; doses(i)=i; End;
 Run;
 Proc Print Data=Dinput; Var Doses1-Doses8 RRo1-RRo8; Run;
%CRM(nSims=1000, nPts=8, nLevels=8, b=150, aMin=0, aMax=3,
MTRate=0.17);
%CRM(nSims=1000, nPts=16, nLevels=8, b=150, aMin=0, aMax=3,
MTRate=0.17);
Run;
```
‹‹SAS‹‹

Table 15.2: Adaptive Dose-Response Simulation Results

Method	Mean N	Mean DLTs	Mean MDT	SdMTD
TER	18.4	1.47	5.33	1.82
STER	19.2	1.60	5.27	1.77
CRM	8.0	1.24	6.01	0.07
CRM	16.0	2.65	6.00	0.06

In CRM, the following logistic model is used

$$p = \frac{1}{1 + 150\exp(-a\,i)},$$

where i = dose level (we can also use actual dose) and the prior distribution for parameter a is flat over [0, 3].

Note the true MTD is dose level 6 ($228\ mg/m^2$). The simulation results are summarized in Table 15.2. The average predicted MTD (dose level)

is 5.33 with TER and 5.27 with STER, which are underestimated. In contrast, the average MTD for the two CRM with sample-size 8 and 16 patients accurately predict the true MTD. From precision point of view, even with a smaller sample-size, the standard deviation of MTD (SdMTD) is much smaller for CRM than both TER and STER. It can be seen that increasing sample-size in CRM may not materially increase the precision, but could increase the DLTs or responses, which is not desirable. In the current scenario, the CRM design with 8 patients is the best among the four designs. However, we should be aware that the performance of CRM is dependent on the goodness of the model specification.

15.3 Summary and Discussion

We have studied the traditional algorithm-based and model-based approaches for dose-response trials. The efficiencies of the approaches are dependent on several aspects, such as the situation in which the next patient is enrolled before we have the response data from the previous patient. In this case, the efficacy of the TER and CRM may be reduced. There may also be a limit for dose escape, which may also reduce the efficiency of the CRM.

In addition to $A+B$ escalation algorithms, many other algorithms have been proposed (Chevret, 2006 and Ting, 2006). For example, Shih and Lin (2006) modified $A+B$ and derived closed form solutions for the modified algorithms. They were motivated by the following: (1) in a traditional $A+B$ design, the previous dose level is always declared the maximum tolerated dose (MTD), and the current dose has no chance at all of being declared the MTD; (2) the starting dose cannot necessarily be the lowest dose; and (3) the design may be a two-stage escalation design used to accelerate the escalation in the early part of the trial with one or two patients per dose level. However, the values from (1) and (2) are questionable because any dose can be the current or previous dose level, and just like dose jump, dosing the first patient at a higher dose level may put certain individuals at risk, even though the overall toxicity (e.g., number of DLTs) may be reduced.

CRM can be used with other monotonic or non-monotonic models and can be combined with response-adaptive randomization or a drop-loser design (Chang, et al., 2005).

Problem

15.1 Modify SAS Macro SAS 15.1 using $m + n$ dose-escalation rules.

15.2 Modify SAS Macro SAS 15.2 to include a control arm; then design a two-arm study (control and active) using CRM.

Chapter 16

Bayesian Adaptive Design

16.1 Introduction

The use of adaptive trial designs could greatly improve the efficiency of drug development; the incorporation of the Bayesian approach is one more step further to this direction. Bayesian approaches provide a powerful tool for streamlining sequential learning processes, predicting future results, and synthesizing evidences across different resources. The resulting outcomes using Bayesian approaches are easier to interpret and more informative for decision making.

The Bayesian approach has been widely studied recently in drug development in areas such as clinical trial design (Spiegelhalter, Abrams, and Myles, 2004), pharmacovigilance(Hauben, et al., 2005; Hauben and Reich, 2004) and pharmacoeconomics (Ades, et al., 2006; Iglesias and Claxton, 2006). Goodman (2005), Louis (2005) and Berry (2005) give excellent introductions to using Bayesian methods in clinical trials and discuss relevant issues.

There are two different ways of using the Bayesian approach in clinical trials: the frequentist-Bayesian hybrid approach and the pure Bayesian approach. In the frequentist-Bayesian hybrid approach, the frequentist statistical criterion for efficacy claim is used, i.e., p-value $< \alpha$, but the Bayesian approach is used to achieve better design or decision-making. The pure Bayesian approach would suggest a change in the current significance criterion to use the so-called Bayesian significance or other Bayesian criteria (Section 16.3).

The Bayesian approach can be used to determine the best strategy available at the time. It can be used to monitor trials, predict outcomes, anticipate problems, and suggest early remedies. The Bayesian approach in combination with adaptive designs (Section 8) will allow for configuring the best strategy over time.

In this chapter, a comparative study of frequentist and Bayesian approaches is pursued to objectively evaluate the two different approaches. The differences are identified in various aspects such as learning mechanism, trial design, monitoring, data analysis, and result interpretation. The regulatory aspects of Bayesian approaches are reviewed, challenges in using Bayesian designs are addressed, and steps for planning a Bayesian trial are outlined.

16.2 Bayesian Learning Mechanism

We acquire knowledge through a sequence of learning processes. We form a perception about a certain thing based on prior experiences, i.e., prior knowledge. This knowledge is updated when new facts are observed. This learning mechanism is the central idea of the Bayesian approach. Bayes' theorem, the foundation of the Bayesian approach, can be expressed as: Posterior distribution = C(Likelihood• Prior distribution), where C is a normalization constant that can be calculated. The posterior distribution that is used to make inferences about the treatment effect is a combination of evidences from both prior (or historical information sources) and the current trials. Note that prior refers to the knowledge before the current trial and posterior distribution is the knowledge after the current trial.

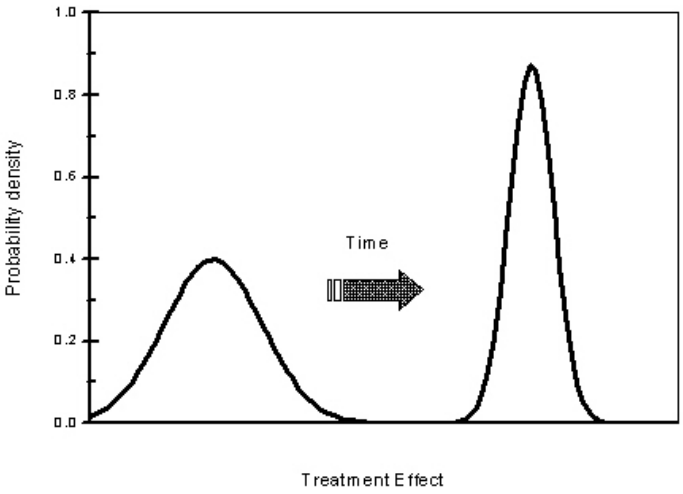


Figure 16.1: Bayesian Learning Process

Bayes' rule reveals the important relationships among prior knowledge, new evidence, and updated knowledge (posterior probability). It reflects the human learning process. Whether you are a statistician, a physician, or a regulatory agent, you are using the Bayesian approach constantly, formally or informally, consciously or unconsciously.

The differences between the frequentist and Bayesian approaches are reflected in the learning process. Both frequentist and Bayesian approaches use prior probabilities, but in quite different ways. The frequentist approach uses prior information segmentally, while the Bayesian approach uses the prior rigorously and intelligently. The frequentist approach considers a population parameter to be a fixed number; the Bayesian approach views the population parameter as a random variable with a probability distribution, which can be updated as more information is accumulated. The Bayesian approach can be used to gain knowledge about treatment effect over time; this is illustrated in Figure 16.1. We can see that the uncertainty about the treatment effect is reduced over time.

Drug development is a process that switches between statistics and probability. Statistics is the study of (drawing a conclusion about) population characteristics based on the observed sample; probability is the study of the sample properties based on the characteristics of the population. For example, to design a phase II trial, we estimate the treatment effect from prior information and use that to calculate the sample-size and the probability of success (power); at the end of the trial, we estimate the treatment effect again based on the phase II trial results, then use that to design the next trial (phase III), and predict power. At the end of the phase III trial, we estimate the treatment effect again. However, the frequentist and Bayesian approaches make the inference and calculate probability differently. The frequentist approach assumes a single known treatment effect at each phase in drug development and calculates sample-size based on that single number; the Bayesian approach, on the other hand, realistically uses the posterior distribution of the treatment effect at the end of each phase for sample-size calculation for the next phase. The Bayesian approach also allows for knowledge to be updated when new information becomes available and uses it during the drug development process.

16.3 Bayesian Basics

16.3.1 *Bayes' Rule*

Denote prior distribution $\pi(\theta)$, and the sample distribution $f(x|\theta)$. As mentioned in Chapter 15, the following are four basic Bayesian elements:

(a) the joint distribution of (θ, x) given by

$$\varphi(\theta, x) = f(x|\theta)\,\pi(\theta), \tag{16.1}$$

(b) the marginal distribution of x given by

$$m(x) = \int \varphi(\theta, x)\,d\theta = \int f(x|\theta)\,\pi(\theta)\,d\theta, \tag{16.2}$$

(c) the posterior distribution of θ given by Bayes' formula

$$\pi(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{m(x)}, \text{ and} \tag{16.3}$$

(d) the predictive probability distribution given by

$$P(y|x) = \int P(x|y,\theta)\,\pi(\theta|x)\,d\theta. \tag{16.4}$$

Example 16.1 Beta Posterior Distribution
Assume that $X \sim Bin(n,p)$ and $p \sim Beta(\alpha,\beta)$.
The sample distribution is given by

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \; x = 0, 1, ..., n. \tag{16.5}$$

The prior about the parameter p is given by

$$\pi(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \; 0 \le p \le 1, \tag{16.6}$$

where beta function $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.
The joint distribution then is given by

$$\varphi(p,x) = \frac{\binom{n}{x}}{B(\alpha,\beta)} p^{\alpha+x-1}(1-p)^{n-x+\beta-1} \tag{16.7}$$

and the marginal distribution is

$$m(x) = \frac{\binom{n}{x}}{B(\alpha,\beta)} B(\alpha+x, n-x+\beta). \tag{16.8}$$

Therefore the posterior distribution is given by

$$\pi\left(p|x\right)=\frac{p^{\alpha+x-1}\left(1-p\right)^{n-x+\beta-1}}{B\left(\alpha+x,\beta+n-x\right)}=Beta\left(\alpha+x,\beta+n-x\right). \tag{16.9}$$

Example 16.2 Normal Posterior Distribution
Assume that $X \sim N\left(\theta,\sigma^2/n\right)$ and $\theta \sim N\left(\mu,\sigma^2/n_0\right)$.
The posterior distribution can be written as

$$\pi\left(\theta|X\right)\propto f\left(X|\theta\right)\pi\left(\theta\right) \tag{16.10}$$

or

$$\pi\left(\theta|X\right)=Ce^{-\frac{(X-\theta)^2 n}{2\sigma^2}}e^{-\frac{(\theta-\mu)^2 n_0}{2\sigma^2}}, \tag{16.11}$$

where C is a constant.

We immediately recognize that (16.11) is the normal distribution of $N\left(\frac{n_0\mu+nX}{n_0+n},\frac{\sigma^2}{n_0+n}\right)$.

We now wish to make predictions concerning future values of X, taking into account our uncertainty about its mean θ. We may write $X=(X-\theta)+\theta$, and so can consider X as being the sum of two independent quantities: $(X-\theta)\sim N\left(0,\sigma^2/n\right)$, and $\theta\sim N\left(\mu,\sigma^2/n_0\right)$. The predictive probability distribution is given by (Spiegelhalter, et al., 2004),

$$X\sim N\left(\mu,\sigma^2\left(\frac{1}{n}+\frac{1}{n_0}\right)\right). \tag{16.12}$$

If we have already observed x_{n_1}, the mean of the first n_1 observations, the predictive distribution is given by

$$X|x_{n_1}\sim N\left(\frac{n_0\mu+n_1x_{n_1}}{n_0+n_1},\sigma^2\left(\frac{1}{n_0+n_1}+\frac{1}{n}\right)\right). \tag{16.13}$$

16.3.2 *Conjugate Family of Distributions*

A family F of probability distribution on Θ is said to be conjugate (or closed under sampling) if, or every $\pi\in F$, the posterior distribution $\pi\left(\theta|x\right)$ also belongs to F.

The main interest of conjugacy becomes more apparent when F is as small as possible and parameterized. When F is parameterized, switching from prior to posterior distribution is reduced to an updating of the corresponding parameters. This is a main reason why conjugate priors are so

popular, as the posterior distributions are always computable, at least to a certain extent.

The conjugate prior approach, which originated in Raiffa and Schlaifer (1961), can be partially justified through an invariance reasoning. Updating the model should not be radical, e.g., only the values of the parameters not the model or function itself is updated. Commonly used conjugate families are presented in Tables 16.1 and 16.2.

Table 16.1: Commonly Used Conjugate Families

Model $f(x\|\theta)$	Prior $\pi(\theta)$	Posterior $\pi(\theta\|x)$
Normal $N(\theta, \sigma^2)$	$N(\mu, \tau^2)$	$N\left(\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2},\right)$
Poisson $P(\theta)$	$G(\alpha, \beta)$	$G(\alpha + x, \beta + 1)$
Gamma $G(\nu, \theta)$	$G(\alpha, \beta)$	$G(\alpha + \nu, \beta + x)$
Binomial $Bin(n, \theta)$	$Beta(\alpha, \beta)$	$Beta(\alpha + x, \beta + n - x)$
Neg. Bin $NB(m, \theta)$	$Beta(\alpha, \beta)$	$Beta(\alpha + m, \beta + x)$

For conjugate family distributions, the estimations can easily be obtained and are summarized in Table 16.2.

Table 16.2: Estimation of Conjugate Families

Distribution	Conjugate distribution	Posterior expectation
Normal $N(\theta, \sigma^2)$	$N(\mu, \tau^2)$	$\frac{\mu\sigma^2+\tau^2 x}{\sigma^2+\tau^2}$
Poisson $P(\theta)$	$G(\alpha, \beta)$	$\frac{\alpha+x}{\beta+1}$
Gamma $G(\nu, \theta)$	$G(\alpha, \beta)$	$\frac{\alpha+\nu}{\beta+x}$
Binomial $Bin(n, \theta)$	$Beta(\alpha, \beta)$	$\frac{\alpha+x}{\alpha+\beta+n}$
Neg. Bin $NB(m, \theta)$	$Beta(\alpha, \beta)$	$\frac{\alpha+n}{\alpha+\beta+x+n}$

16.4 Trial Design

16.4.1 *Bayesian for Classic Design*

We are going to use an example to illustrate some differences between Bayesian and frequentist approaches in trial design.

Example 16.3 Prior Effect on Power

Consider a two-arm parallel design comparing a test treatment with a control. Suppose that, based on published data from 3 clinical trials of similar size, the prior probabilities for effect size are 0.1, 0.25, and 0.4 with 1/3 probability for each.

For the 2-arm trial, the power is a function of effect size of ε, i.e,

$$\text{power}\,(\varepsilon) = \Phi\left(\frac{\sqrt{n}\varepsilon}{2} - z_{1-\alpha}\right), \tag{16.14}$$

where Φ is c.d.f. of the standard normal distribution.

Considering the uncertainty of ε, i.e., prior $\pi(\varepsilon)$, the expected power

$$\mathrm{P}_{\text{exp}} = \int \Phi\left(\frac{\sqrt{n}\varepsilon}{2} - z_{1-\alpha}\right)\pi(\varepsilon)\,d\varepsilon. \tag{16.15}$$

A numerical integration is usually required for evaluation (16.15).

To illustrate the implication of (16.15), let's assume one-sided $\alpha = 0.025$, $z_{1-\alpha} = 1.96$, the prior

$$\pi(\varepsilon) = \begin{cases} 1/3, & \varepsilon = 0.1, 0.25, 0.4 \\ 0, & \text{otherwise.} \end{cases} \tag{16.16}$$

Conventionally we use the mean (median) of the effect size $\bar{\varepsilon} = 0.25$ to design the trial and calculate the sample-size. For the two-arm balanced design with $\beta = 0.2$ or *power* $= 80\%$, using classic approach, the total sample is given by

$$n = \frac{4(z_{1-a} + z_{1-\beta})^2}{\varepsilon^2} = \frac{4\,(1.96 + 0.842)^2}{0.25^2} = 502. \tag{16.17}$$

However, if Bayesian approach is used, the expected power from (16.15) is

$$\begin{aligned} &\mathrm{P}_{\text{exp}} \\ &= \frac{1}{3}\left[\Phi\left(\frac{0.1\sqrt{n}}{2} - z_{1-\alpha}\right) + \Phi\left(\frac{0.25\sqrt{n}}{2} - z_{1-\alpha}\right) + \Phi\left(\frac{0.4\sqrt{n}}{2} - z_{1-\alpha}\right)\right] \\ &= \frac{1}{3}\left[\Phi\left(-0.839\,73\right) + \Phi\left(0.840\,67\right) + \Phi\left(2.521\,1\right)\right] \\ &= \frac{1}{3}\left(0.2005 + 0.7997 + 0.9942\right) = 0.664\,8 = 66\%. \end{aligned} \tag{16.18}$$

We can see that the expected power is only 66%, therefore, we should increase sample-size.

With the Bayesian approach that considers the uncertainty of the effect size, the expected power with a sample-size of 252 is the average of the three powers calculated using the 3 different effect sizes (0.1, 0.25, and 0.4), which turns out to be 66%, much lower than 80% as the frequentist

approach claimed. Therefore, to reach the desired power, it is necessary to increase the sample-size.

This is an example of a Bayesian-frequentist hybrid approach, i.e., the Bayesian approach is used for the trial design to increase the probability of success given the final statistical criterion being p-value $\leq \alpha = 0.025$.

Example 16.4 Power with Normal Prior

If the prior $\pi(\varepsilon) = N\left(\mu, \sigma^2/n_0\right)$, then the expected power can be obtained using the predictive distribution (16.12) and evaluating the chance of the critical event $\left(X > \frac{1}{\sqrt{n}} z_{1-\alpha}\sigma\right)$ occurring, which is given by

$$P_{\text{exp}} = \Phi\left(\sqrt{\frac{n_0}{n_0+n}}\left(\frac{\mu\sqrt{n}}{\sigma} - z_{1-\alpha}\right)\right). \tag{16.19}$$

Now let's look at the expected total example size. The total sample-size is function of the effect size ε, i.e.,

$$n(\varepsilon) = \frac{4(z_{1-a} + z_{1-\beta})^2}{\varepsilon^2}. \tag{16.20}$$

Therefore, the expected total sample-size is given by

$$n_{\text{exp}} = \int \frac{4(z_{1-a} + z_{1-\beta})^2}{\varepsilon^2}\pi(\varepsilon)\, d\varepsilon. \tag{16.21}$$

For flat prior $\pi(\varepsilon) \sim \frac{1}{b-a}, [a \leq \varepsilon \leq b]$,

$$\begin{aligned} n_{\text{exp}} &= \int_a^b \frac{4(z_{1-a} + z_{1-\beta})^2}{\varepsilon^2}\frac{1}{b-a} d\varepsilon \\ &= \frac{4}{ab}(z_{1-a} + z_{1-\beta})^2. \end{aligned} \tag{16.22}$$

The sample-size ratio $R_n = \frac{n_{\text{exp}}}{n} = \frac{\varepsilon^2}{ab}$. For example $\varepsilon = 0.25, \alpha = 0.025, \beta = 0.8, n = 502, a = 0.1, b = 0.4$ (note that $(a+b)/2 = \varepsilon$), $R_n = \frac{0.25^2}{(0.1)(0.4)} = 1.56$. It indicates again that the frequentist approach could substantially underestimate the sample-size required for achieving the target power.

16.4.2 *Bayesian Power*

To test the null hypothesis $\theta \leq 0$ against an alternative hypothesis $\theta > 0$. Bayesian significance is defined as $P^B = P(\theta < 0|data) < \alpha_B$. Using the posterior distribution, the Bayesian significance can be easily found.

Example 16.5 Bayesian Power

For normal distribution prior and data, the posterior distribution is given by

$$\pi(\theta|x) = N\left(\frac{n_0\mu_0 + n\bar{x}}{n_0 + n}, \frac{\sigma^2}{n_0 + n}\right). \tag{16.23}$$

Bayesian significance is claimed if the parameter estimate $\bar{x}$ satisfies

$$\bar{x} > \frac{\sqrt{n_0 + n}z_{1-\alpha_B}\sigma - n_0\mu_0}{n}, \tag{16.24}$$

$$\frac{\sqrt{n_0 + n}z_{1-\alpha_B}\sigma - n_0\mu_0}{n}\frac{\sqrt{n}}{\sigma} + \mu.$$

Therefore the Bayesian power is then given by

$$P_B(n) = 1 - \Phi\left(z_{1-\alpha_B}\sqrt{\frac{n_0}{n} + 1} - n_0\sqrt{n}\frac{\mu_0}{\sigma} + \mu\right), \tag{16.25}$$

where μ is the true mean for the population.

Example 16.6 Trial Design Using Bayesian Power

Suppose in a phase II two-arm hypotension study with the SBP reduction as the primary endpoint, the estimated treatment effect is normal distribution, i.e., $\theta \sim N\left(\mu, \frac{2\sigma^2}{n_0}\right)$. The trial is designed with a Bayesian power of $(1-\beta_B)$ at the Bayesian significance level $\alpha_B = 0.2$. For the sample-size, the sample mean difference can be expressed as $\hat{\theta} \sim N\left(\theta, \frac{2\sigma^2}{n}\right)$, where $n =$ sample-size per group. For a large sample-size, we can assume that $\hat{\sigma}$ is constant. Therefore the sample-size n is the solution for the following

$$1 - \Phi\left(z_{1-\alpha_B}\sqrt{\frac{n_0}{n} + 1} - n_0\sqrt{n}\frac{\mu_0}{\sigma} + \mu\right) = 1 - \beta_B. \tag{16.26}$$

That is,

$$z_{1-\alpha_B}\sqrt{\frac{n_0}{n}+1}-n_0\sqrt{n}\frac{\mu_0}{\sigma}+\mu=z_{\beta_B}. \tag{16.27}$$

Equation (16.27) can be solved numerically for sample-size n.

16.4.3 *Frequentist Optimization*

Simon's Two-Stage Design

Simon's two-stage optimal design (Simon, 1989) is a commonly used design for single-arm oncology trials. The hypothesis testing can be stated as

$$H_o: R < R_1 \text{ vs. } H_a: R \geq R_2,$$

where R is response rate.

The trial has one interim analysis (IA). At IA, if the observed $\hat{R} < R_0$ or the number of response is less than a constant n_{1c}, then stop the trial for futility. Otherwise trial continues. At the final, if the total number of responses is larger than or equal to n_c, then reject H_o. Otherwise, accept H_o. For a given α and power, different values of n_{1c}, and n_c will lead to different maximum sample-size and expected sample sizes under H_o and H_a. The optimal design minimizes the expected sample-size under H_o and the MinMax design minimizes the maximum sample-size.

SAS Macro 16.1 can be used for a single-arm two-stage design with interim futility stopping. For a given constant n, the macro will search the best set of designs with sample-size ranging from $0.7n$ to $1.5n$. The SAS variables are defined as follows: **Alpha0** = target one-sided significance level, **Alpha** = actural one-sided alpha, **po** and **pa** = response rates under H_o and H_a, respectively, **n** = sample-size group, **n1** = sample-size at interim analysis for early futility stopping, **n1c** = critical value for futility: If the number of responses at the first stage is less than **n1c**, stop for futility; otherwise, continue to the second stage. If the total number of responses from the two stages >= **nc**, claim efficacy. **PrEFSHo** and **PrEFSHa** are the probabilities of early stopping under H_o and H_a, respectively. **ExpNo** and **ExpNa** are the expected sample sizes under H_o and H_a, respectively.

>>SAS Macro 16.1: Simon Two-Stage Futility Design>>

```
%Macro TwoStageDesign(n=50, po=0.15, pa=0.3, n1=20, n1c=2,
alpha0=0.1);
Data TwoStageBin;
 retain alpha power;
```

```
drop i p1o p2o p1a p2a n2;
n1=&n1; n1c=&n1c; po=&po; pa=&pa; * Remove "&".;
do n=round(0.7*&n) to round(1.5*&n);
     n2=n-&n1;
    do nc=n1c to n;
         alpha=0;
         power=0;
         do i=max(n1c,nc-n2) to n1;
              p1o=ProbBnml(po, n1,i)-ProbBnml(po, n1,i-1);
             p2o=1;
             if nc-i>0 then p2o=1-ProbBnml(po, n2,nc-i-1);
             alpha=alpha+p1o*p2o;
             p1a=ProbBnml(pa, n1,i)-ProbBnml(pa, n1,i-1);
             p2a=1;
             if nc-i>0 then p2a=1-ProbBnml(pa, n2,nc-i-1);
             power=power+p1a*p2a;
         end;
         if alpha>0.8*&alpha0 && alpha<1.2*&alpha0 then do;
             PrEFSHo=ProbBnml(po, n1,n1c-1);
             ExpNo=PrEFSHo*n1+(1-PrEFSHo)*n;
             PrEFSHa=ProbBnml(pa, n1,n1c-1);
             ExpNa=PrEFSHa*n1+(1-PrEFSHa)*n;
             output;
         end;
    end;
end;
run;
proc print; run;
run;
%Mend TwoStageDesign;
```

‹‹SAS‹‹

Example 16.7 Simon Two-Stage Optimal Design

Suppose we design a single-arm, phase-II oncology trial using Simon's two-stage optimal design. The response rates are assumed 0.05 and 0.25 under the null hypothesis and the alternative hypothesis, respectively. For one-sided $\alpha = 0.05$ and power = 80%, the sample-size at stage 1 is $n_1 = 9$. The cumulative sample-size at stage 2 is $n = 17$. The actual overall $\alpha = 0.047$, the actual power = 0.812. The stopping rules are specified as follows: At stage 1, stop and accept the null hypothesis if the response rate is less than 1/9. Otherwise, continue to stage 2. The probability of stopping for

futility is 0.63 when H_o is true and 0.075 when H_a is true. At stage 2, accept the null hypothesis if the response rate is less than or equal to 2/17. Otherwise, reject the null hypothesis.

The results can be generated using ExpDesign Studio®, or the following SAS macro call:

```
>>SAS>>
%TwoStageDesign(n=17, po=0.05, pa=0.25, n1=9, n1c=1, alpha0=0.05);
<<SAS<<
```

16.4.4 *Bayesian Optimal Adaptive Designs*

As discussed earlier in Section 12.4, Bayesian decision theory can be used to optimize trial designs. The Bayesian approach is decision-oriented. A Bayesian views statistical inference as a problem in belief dynamics, or use of evidence about a phenomenon to revise our knowledge about it. In distinguishing from a frequentist, for a Bayesian, statistical inference cannot be treated entirely independently of the context of the decisions that will be made on the basis of the inferences.

There are many different scenarios of reality with associated probabilities (prior distribution) and many possible adaptive designs with associated probabilistic outcomes (good and bad). Evaluation criterion can be the utility index that can be the aggregation of overall patients' health outcomes. Bayesian optimal design is to achieve the maximum expected utility under financial, time, and other constraints. We will use two-arm designs to illustrate the approach. The three designs we are going to compare are: classic approach with a two-arm phase II trial followed by a two-arm phase III trial, and two different group sequential designs (seamless designs).

For each design, calculate the utility and weighted by its prior probability to obtain the expected utility for the design. The optimal design is the one with maximum expected utility.

Example 16.8 Bayesian Optimal Design

Suppose prior knowledge about treatment effect is determined as shown in Table 16.3. We are going to compare the classic design and Bayesian adaptive designs.

Table 16.3: Prior Knowledge

Scenario	Effect size	Prior prob.
1	0	0.2
2	0.1	0.2
3	0.2	0.6

The Classic Design: Assume there is no dose selection issue. For classic design, we use a phase-II and a phase-III trial (assume that just one phase-III trial is required for approval). For the phase II trial, we assume $\delta = 0.2$, one-sided $\alpha = 0.1$ and power = 0.8, the total sample-size required is $n_1 = 450$. For the phase-III trial, assume $\delta = 0.14$ (calculated by 0.2(0) + 0.2(0.1) + 0.6(0.2) from Table 16.3), one-sided $\alpha = 0.025$, power = 0.9, the total sample-size required is $n = 2144$. If the phase II didn't show statistical significance, we will not conduct the phase-III trial (In practice, the rule is not always followed). The probability of continuing the trial at phase-II is the weighted continual probability in Table 16.4, i.e.,

$$P_c = \Sigma_{i=1}^{3} P_c(i)\,\pi(i) = 0.2(0.1) + 0.2(0.4) + 0.6(0.8) = 0.58.$$

Therefore, the expected sample-size for phase-II and III trials together is

$$\bar{N} = n_1 + P_c n = 450 + 0.58(2144) = 1780.$$

The expected overall power is given by

$$\begin{aligned}\bar{P} &= \Sigma_{i=1}^{3} P_c(i)\,\pi(i)\,P_3(i) \\ &= (0.2)(0.1)(0.025) + (0.2)(0.4)(0.639) + (0.6)(0.8)(0.996) \\ &= 0.53.\end{aligned}$$

Table 16.4: Characteristics of Classic Phase II and III Designs

			Prob. of continue	Phase III
Scenario, i	Effect size	Prior prob. π	to Phase III, P_c	Power, P_3
1	0	0.2	0.1	0.025
2	0.1	0.2	0.4	0.639
3	0.2	0.6	0.8	0.996

In conclusion, the classic phase II trial followed by a phase III trial has overall power = 53% with expected grand combined sample = 1780.

Seamless Design With OB-F Boundary: Use one-sided $\alpha = 0.025$, power =0.90, with O'Brien-Fleming efficacy stopping boundary and symmetrical futility stopping boundary. One interim analysis will be conducted

when 50% patients are enrolled. We can use the ExpDesign Studio® (Figure 16.2) or SAS Macro 6.1 to simulate the trial. The operating characteristics of the design are summarized in Table 16.5.

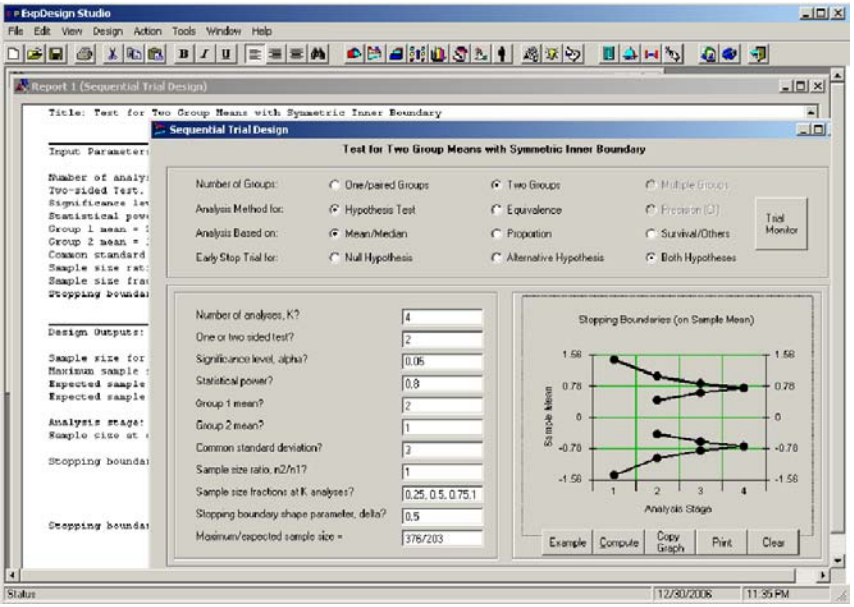


Figure 16.2: ExpDesign Studio

Table 16.5: Characteristics of Seamless Design (OBF)

Scenario, i	Effect size	Prior prob. π	N_{exp}	Power
1	0	0.2	1600	0.025
2	0.1	0.2	1712	0.46
3	0.2	0.6	1186	0.98

Average example size can be calculated:

$$\begin{aligned} N_{\text{exp}} &= \Sigma\pi\left(i\right) N_{\text{exp}}\left(i\right) \\ &= 0.2(1600) + 0.2(1712) + 0.6(1186) \\ &= 1374. \end{aligned}$$

Average power can be calculated:

$$\begin{aligned} P_{\text{exp}} &= \Sigma\pi\left(i\right) N_{\text{exp}}\left(i\right) Power\left(i\right) \\ &= 0.2(0.025) + 0.2(0.46) + 0.6(0.98) \\ &= 0.69 \end{aligned}$$

Seamless Design with Pocock Boundary: We now use Pocock boundary efficacy stopping boundary and the symmetric futility stopping boundary to design the trial. The operating characteristics are summarized in Table 16.6.

Table 16.6: Characteristics of Seamless Design (Pocock)

Scenario, i	Effect size	Prior prob. π	N_{exp}	Power
1	0	0.2	1492	0.025
2	0.1	0.2	1856	0.64
3	0.2	0.6	1368	0.996

Average sample-size is given by

$$\begin{aligned} N_{\text{exp}} &= \Sigma\pi\left(i\right) N_{\text{exp}}\left(i\right) \\ &= 0.2(1492) + 0.2(1856) + 0.6(1368) \\ &= 1490. \end{aligned}$$

Average power is given by

$$\begin{aligned} P_{\text{exp}} &= \Sigma\pi\left(i\right) N_{\text{exp}}\left(i\right) Power\left(i\right) \\ &= 0.2(0.025) + 0.2(0.64) + 0.6(0.996) \\ &= 0.73. \end{aligned}$$

Let's compare the different designs from a financial perspective. Assume per-patient cost in the trial = \$50k, dollar value of approval before deducting the trial cost = \$1B. Time savings are not included in the calculation. Therefore, the expected utility can be expressed as

$$\text{Expected utility} = (\text{Average power})(\$80\text{M})\text{-}(\text{N}_{\text{exp}})(\$50\text{K}).$$

The resulting expected utility for the three designs are summarized in Table 16.7. We can see that the Pocock design is the best among the three designs based on both power and the expected utility.

Table 16.7: Comparison of Classic and Seamless Designs

Design	N_{max}	Average N_{exp}	Average power	Expected utility
Classic		1500	0.59	\$0.515B
BOF		1374	0.69	\$0.621B
Pocock		1490	0.73	\$0.656B

16.5 Trial Monitoring

Conditional and Predictive Powers

The conditional and predictive power can be used to monitor adaptive designs.

Assume X has binomial distribution. Given x out of n_1 patients have the response at first stage, what is the probability (conditional power) of having at least y additional responses out of n_2 additional patients at second stage? The conditional power can be obtained using the frequentist approach:

$$P\left(y|x,n_1,n_2\right)=\sum_{i=y}^{n_2}\left(\begin{array}{c} n_2 \\ i \end{array}\right)\left(\frac{x}{n_1}\right)^i(1-\frac{x}{n_1})^{n_2-i}. \tag{16.28}$$

Now let's look at Bayesian point of view. For the nonformative prior i.e., p is uniformly distributed in $[0,1]$, we have

$$P(X=x|p)=\binom{n_1}{x}p^x(1-p)^{n_1-x},$$

$$P(a<p<b\ \cap\ X=x)=\int_a^b\binom{n_1}{x}p^x(1-p)^{n_1-x}dp,$$

$$P(X=x)=\int_0^1\binom{n_1}{x}p^x(1-p)^{n_1-x}dp,$$

and

$$\begin{aligned}P(a<p<b\mid X=x)&=\frac{\int_a^b\binom{n_1}{x}p^x(1-p)^{n_1-x}dp}{\int_0^1\binom{n_1}{x}p^x(1-p)^{n_1-x}dp}\\&=\frac{\int_a^b p^x(1-p)^{n_1-x}dp}{B(x+1,n_1-x+1)},\end{aligned}$$

where *Beta* function

$$B(x+1, n_1 - x + 1) = \frac{\Gamma(x+1)\Gamma(n_1 - x + 1)}{\Gamma(n_1 + 2)}.$$

Therefore the posterior distribution of p is a *Beta* distribution

$$\pi(p|x) = \frac{p^x(1-p)^{n_1-x}}{B(x+1, n_1 - x + 1)}. \tag{16.29}$$

The predictive power (to differentiate from frequentist's conditional power) or the predictive probability of having at least y responders out of additional n_2 patients is given by

$$\begin{aligned} P(y|x, n_1, n_2) &= \int_0^1 P(X \geq y|p, n_2)\pi(p|x)\, dp \\ &= \int_0^1 \sum_{i=y}^{n_2} \binom{n_2}{i} p^i (1-p)^{n_2-i} \frac{p^x(1-p)^{n_1-x}}{B(x+1, n_1 - x + 1)} dp. \end{aligned}$$

Carrying out the integration, we have

$$P(y|x, n_1, n_2) = \sum_{i=y}^{n_2} \binom{n_2}{i} \frac{B(x+i+1, n_2 + n_1 - x - i + 1)}{B(x+1, n_1 - x + 1)}, \tag{16.30}$$

where we have used the results:

$$\int_0^1 p^a (1-p)^b \, dp = B(a+1, b+1).$$

16.6 Analysis of Data

An important feature of the Bayesian approach is that it provides an ideal methodology for synthesizing information. This is often accomplished under the so-called exchangeability assumption. As pointed out by the Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials (FDA, 2006), exchangeability is a key idea in statistical inference in general, but it is particularly important in the Bayesian approach. If the patients in a clinical trial are exchangeable with the target population, then the clinical trial can be used to make inferences about the entire population. Data from different trials are more or less correlated. Exchangeability is a way

to characterize the correlation. Suppose we are interested in making inferences on many parameters θ_1, θ_2,...θ_k measured on K 'units,' which may, for example, be true treatment effects in subpopulations, investigative centers, or a sequence of trials. The parameters θ_1, θ_2,...θ_k are exchangeable, if the labels (subscripts) are not systematically associated with the values of the parameters. In other words, θ_1, θ_2,...θ_k are samples from the same distribution of some "super population." It is important to know that two trials may be exchangeable only after adjustments are made for other confounding factors with the appropriate statistical model (Cornfield, 1976).

In the frequentist approach, the parameter θ is considered a fixed value; hence, a fixed effect model is used for the analysis. However, in the Bayesian paradigm, the parameter is considered a random variable (hence, θ_1, θ_2,...θ_k above can be considered a random sample from a common super population). Therefore, the random-effect model with the exchangeability assumption is used for the analysis. This relationship between subpopulations and super populations can exist on multiple levels — a hierarchical model. Because of the similarity or exchangeability, data can be pooled across patient and disease groups within the same trial, and across trials, using a hierarchical model to obtain more precise estimates.

An example of a hierarchical model is illustrated in the Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials: (Cornfield, 1976). Suppose you want to combine information from a treatment registry of an approved device with results from a new study. A model of two hierarchical levels (the patient level and the study level) is used. In the first (patient) level of the hierarchy, exchangeability is assumed for each of the studies. However, registry patients are not exchangeable with patients in the current study, so patient data from the registry and the current study cannot simply be pooled. The second (study) level of the hierarchy applies a model that assumes that the success probabilities from the registry and the current study are exchangeable after adjustment for covariates. Due to the use of this hierarchical model, the registry provides some information about the success probability for the current study, although not as much information as if the patients in the two groups were pooled directly as in a homogeneous case.

To give a real-life example of the Bayesian approach, in 2003, the FDA approved a drug that combines pravastatin, a cholesterol-lowering agent, with aspirin, based on the use of an exclusively Bayesian analysis of efficacy. The Bayesian approach was used to synthesize information through a meta-analysis of data from five previous pravastatin secondary prevention trials. Hierarchical modeling allowed for diverse sets of patients within the various trials (Berry, 2005).

In contrast, the frequentist approach does not allow for information synergy from different sources (trials) in general, at least in the current regulatory setting.

To illustrate the synergy of evidences, let's consider a very simple example. Suppose we have completed phase II and phase III asthma studies that have similar patient populations, similar treatment regimens, etc. The mean difference in efficacy was 7% improvement from baseline FEV1 with a standard deviation of 22% (Standard error = 0.022, n = 200 per group) in the phase II trial, and 9% improvement with a standard deviation of 20% (Standard error = 0.0126, n = 500 per group) in the phase III trial. If we use phase II as the prior for phase III, we will have the posterior Normal distribution for the percent FEV1 improvement with a mean of 7.5% with a standard deviation 1.1%. We have learned the treatment effect from the phase III data on top of the phase II data, rather than only from the phase III data.

In practice, because of the heterogeneity of the trials, more complicated hierarchical models have to be used to derive the posterior distribution of treatment effect as mentioned earlier for the combined drug approved by FDA in 2003 (Berry, 2005).

16.7 Interpretation of Outcomes

In addition to p-value, confidence interval (CI) is commonly reported as a frequentist outcome of clinical trials. Ironically, frequentist CIs are often misinterpreted as Bayesian credible intervals (BCI) because the concept of CI is difficult to understand and somewhat awkward to interpret. Fortunately, the CI is numerically close to the BCI with a non-informative prior distribution. For this reason, this misunderstanding does not lead to any tragedy.

The concept of a CI is quite difficult for non-statisticians. For example, assume that our population parameter of interest is the population mean. What is the meaning of a 95% CI in this situation? The correct interpretation is based on repeated sampling. If samples of the same size are drawn repeatedly from a population, and a CI is calculated from each sample, then 95% of these intervals should contain the population mean. We can say that the probability of the true mean falling within this set of intervals with various lower and upper bounds is 95% (Figure 16.3). However, we can't say that the probability of the true mean falling within a particularly observed (i.e. fixed) CI is 95%.

The frequentist approach considers the treatment effect to be a fixed

constant and CI bound as random variables, while the Bayesian considers the treatment effect to be a random variable and BCIs as fixed constants when data are given. Another way of saying this is that frequentist inferences are based on the "sample space" (the set of possible outcomes of a trial), while Bayesian inferences are based on the "parameter space" (the posterior distribution). For this reason, BCI is easy to interpret (Figure 16.3). For example, if the 95% BCI of the treatment mean is (1, 3), we can say that there is 95% probability that the true mean falls within (1, 3). However, for a CI of (1, 3), we cannot say anything about the location of the true mean. Clearly, BCI is much more informative than frequentist CI and can be used to design a better trial, as discussed in the following example. Note that these posterior probabilities are completely different from power for the hypothesis testing. Power is the probability of showing statistically that the treatment effect is larger than zero, under the assumption of a certain treatment effect, while Bayesian posterior probability is telling us the probabilities associated with different magnitudes of treatment effect. Power is associated with alpha, while the posterior probability is the updated knowledge about the treatment effect, and is not associated with alpha.

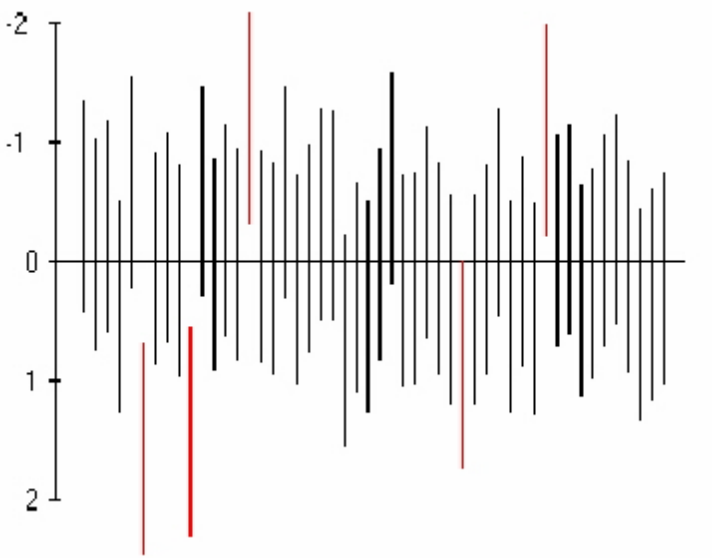


Figure 16.3: Interpretation of Confidence Interval: Five out of 100 intervals do not cover the population mean (0) with $\alpha = 5\%$

16.8 Regulatory Perspective

Bayesian approaches are accepted by the FDA medical device division, CDRH, which recently issued a FDA draft guidance on the use of Bayesian statistics in medical device clinical trials (FDA, 2006). In CDER FDA, although the frequentist approach, i.e., alpha control, is used in the New Drug Application (NDA) approval criteria, information from other sources has never been ignored. In fact, the FDA uses "prior" information in its decision making on clinical trials, which it refers to as informal use of the Bayesian approach.

Janet Woodcock (acting deputy commissioner for operations, FDA) commented at a workshop on Bayesian Clinical trials, "Bayesian approaches to clinical trials are of great interest in the medical product development community because they offer a way to gain valid information in a manner that is potentially more parsimonious of time, resources and investigational subjects than our current methods. The need to streamline the product development process without sacrificing important information has become increasingly apparent." (Woodcock, 2004)

Temple also articulated his view and pointed out that the FDA Evidence Document describes, in qualitative ways, how and when the FDA will take into account other data to reach an effectiveness conclusion based on a single study, and how other controlled trial data can contribute to the present case (FDA, 2005).

The Bayesian philosophy is clearly reflected in the Evidence Document criteria for regulatory approval based on one study (FDA, 2005). The drug carvediol for congestive heart failure can be used to illustrate how prior information plays a role in drug evaluation (Temple, 2005). Carvedilol was already approved for the treatment of congestive heart failure to improve survival and decrease hospitalization. A new study called CAPRICORN studied post-infarction patients with left ventricular dysfunction (ejection fraction less than 40) (Dargie, 2001) to study carveldilol use after a heart attack in people with decreased ejection fractions. The primary endpoint was total mortality (TM). However, in the middle of the study, a new endpoint, death plus cardiovascular hospitalization (DPCH), was added by the data monitoring committee (DMC). TM and DPCH became coprimary endpoints. The multiple adjustments were specified, and the family-wise alpha was split into two parts: 0.005 for TM and 0.045 for DPCH. This meant that to claim efficacy of the drug, either of the following conditions must have been met: (1) p-value for TM less than or equal to 0.005 or (2) p-value for DPCH less than or equal to 0.045. The trial results came out as follows: p-value for DPCH was much larger than 0.045, but p-value

for TM was about 0.03. Based on the modified endpoints, efficacy could not be claimed. However, if the DMC had not changed the endpoint, the efficacy of the drug would obviously have been claimed. So, what to do? After extensive discussions, the FDA agreed with the recommendation by the Cardiorenal Advisory Committee to approve the drug. Temple stated that the fairly explicit reasons were that there were very strong priors here. Carvedilol was unequivocally effective in congestive heart failure; pooled early heart failure trials showed a survival effect for carvedilol and a large study in moderate to severe heart failure (COPERNICUS) showed a clear survival effect (Packer, et al., 2001).

16.9 Summary and Discussions

The Bayesian approach can be used in drug development in two different ways: pure Bayesian and hybrid approaches. The pure Bayesian approach is preferred for virtually all studies before phase III. This is because the Bayesian approach can better incorporate uncertainties at different stages and produce more informative output, such as posterior probability for decision-making and streamlining the process of moving from one phase to the next. For phase III trials, before regulatory agencies accept Bayesian evidence, a Bayesian-frequentist hybrid approach can make better use of the information from earlier phases in the design of phase III studies, as illustrated in the examples above.

Operational challenges are usually similar for Bayesian trials and frequentist trials. However, when Bayesian approaches are used in the clinical development plan and portfolio optimization, they require the ability to rapidly integrate knowledge and experiences from different disciplines into the decision-making process and hence require a shift to a more collaborative working environment.

Problem

16.1 Discuss the frequentist and Bayesian approaches in drug development.

Chapter 17

Planning, Execution, Analysis, and Reporting

17.1 Validity and Integrity

Both the PhRMA white papers (Gallo, et al., 2006) and the discussion paper (Chang and Chow, 2006) emphasize the importance of validity and integrity of adaptive trials. Although controlling the overall type-I error rate at the nominal level (alpha) is essential, it does not imply validity and integrity of the trial. The validity of a trial includes internal and external validities. A study that readily allows its findings to be generalized to the population at large has high external validity. Internal validity refers to the degree to which we are successful in eliminating confounding variables and establishing a cause-effect relationship (treatment effect) within the study itself. There are many different ways in which the internal validity of a study could be jeopardized. Threats to internal validity include instrumentation (case report form, coding shift, evaluation criteria), selection bias (failure in randomization at some level, e.g., a less-sick patient may prefer to wait for enrollment in the confirmatory stage instead of the learning stage, but a sicker patient may not be able to wait), and experimental mortality (informed dropouts). The threats to external validity are protocol amendments, including changes in inclusion/exclusion criteria, which could result in a shift in the target patient population (Chow, Chang, and Pong, 2005), and multiple-endpoints that do not support a common conclusion.

Ensuring integrity is also critical in a seamless design. Integrity means a solid protocol design, excellent execution, unbiased analyses of trial data, and correct interpretation of the results. Integrity means being ethical and avoiding the out-weighting of the risk-benefit ratio of individual patients, trial patients as a whole, and future patients. Integrity also means that regulatory agencies use appropriate approval criteria that balance risk and benefit. The use of a fixed type-I error rate criterion might, in fact, prevent a low-risk and low-benefit drug from being delivered to patients.

17.2 Study Planning

Before the implementation of an adaptive design, it is recommended that the following practical issues be considered. First, determine whether an adaptive design is feasible for the intended trial. For example, will the adaptive design require extraordinary efforts for implementation? Are the level of difficulty and the associated cost justifiable for the gain from the adaptive design? Will the implementation of the adaptive design delay patient recruitment and prolong the duration of the study? Would delayed responses diminish the advantage of the adaptive design? How often will the unblinded analyses be practical, and to whom should the data be unblinded? How should the impact of a Data Monitoring Committee's (DMC's) decisions regarding the trial (e.g., recommending early stopping or other adaptations due to safety concerns) be considered at the design stage? Second, we should ensure the validity of the adaptive design for the intended trial. For example, will the unblinding cause potential bias in treatment assessment? Will the implementation of the adaptive design destroy the randomness? Third, we should have an expectation about the degree of flexibility (sensitivity or robustness). For example, will protocol deviations or violations invalidate the adaptive method? How might an unexpected DMC action affect the power and validity of the design? In designing a trial, we should also consider how to prevent premature release of the interim results to the general public using information masks because releasing information it could affect the rest of the trial and endanger the integrity of the trial. Regarding study design, and we strongly suggest early communication with the regulatory agency and DMC regarding the study protocol and the DMC charter. For broader discussions of planning different adaptive designs, please see the PhRMA full white papers on adaptive design (Quinlan, Gallo, and Krams, 2006; Gallo, 2006, Gaydos, et al., 2006; Maca, et al., 2006; Chuang, et al., 2006)

17.3 Working with Regulatory Agency

Dr. Robert Powell from FDA said that companies should begin a dialogue about adaptive designs with FDA medical officers and statisticians as early as a year before beginning a trial. FDA's Office of Biostatistics Associate Director-Adaptive Design/Pharmacogenomics, Dr. Sue-Jane Wang addressed some of the expectations for adaptive design submissions (Wang, 2006): (1) Is prospectively planned; (2) Has valid statistical approaches on modification of design elements that have alpha control and can be defined

in terms of ICH E-9 standard; (3) Has valid point estimates and confidence interval estimates; (4) Builds on experience from external trials; (5) Takes a "learn" and "confirm" approach; (6) Has standard operating procedures and infrastructure for adaptive process monitoring to avoid bias; (7) Has SOPs on adaptive design decisions; and (8) Includes documentation of actual monitoring process, extent of compliance, and potential effect on study results.

It is important to assure the validity of the adaptive design. Dr. O'Neill said (The Pink Sheet, Dec. 18, 2006, p.24), "We're most concerned about estimating what you think you are estimating," "Is the hypothesis testing appropriate? Do you know what you are rejecting at end of day? Can you say you are controlling the false positive?"

Another reason to communicate with the agency early is that FDA could be of assistance in sharing the data or at least the disease information or models as Dr. Robert Powell indicated. Building off external data and experience is sometimes a crucial element of adaptive design. Just as Pfizer SVP Dr. Declan Doogan pointed out: Drug development is a knowledge-creating business, we have to increase access to data... the FDA is sitting on a pile of wonderful data. How can we access that? We could learn much from other people's failures (The Pink Sheet, Dec. 18, 2006, p.24).

17.4 Trial Monitoring

In practice, it is recognized that there are often deviations from the study protocol when conducting a clinical trial. It is ethical to monitor the trials to ensure that individual subjects are not exposed, or have limited exposure, to unsafe or ineffective treatment regimens. For this purpose, a Data Monitoring Committee (DMC) is usually established. The DMC plays a critical role in monitoring clinical trials. There are common issues that affect a DMC's decision, such as short-term versus long-term treatment effects, early termination philosophies, response to early beneficial trends, response to early unfavorable trends, and response where there are no apparent trends (Ellenberg, Fleming, and DeMets, 2002; Pocock, 2005). It is recommended that a DMC be established to monitor the trial when an adaptive design is employed in clinical trials, especially when many adaptations are considered for allowing greater flexibility.

The stopping rule chosen in the design phase serves as a guideline to a DMC (Ellenberg, Fleming, and DeMets, 2002) as it makes a decision to recommend continuing or stopping a clinical trial. If all aspects of the conduct of the clinical trial adhered exactly to the conditions stipulated

during the design phase, the stopping rule obtained during the design phase could be used directly. However, there are usually complicating factors that must be dealt with during the conduct of the trial.

Deviation of Analysis Schedule: DMC meetings are typically based on the availability of its members, which may be different from the schedules set at the design phase. The enrollment may be different from the assumption made at the design phase. Deviation in the analysis schedule will affect the stopping boundaries; therefore, the boundaries should be recalculated based on the actual schedules.

Deviation of Efficacy Variable Estimation: The true variability of the response variable is never known, but the actual data collected at interim analysis may show that the initial estimates in the design phase are inaccurate. Deviation in the variability could affect the stopping boundaries. In this case, we may want to know the likelihood of success of the trial based on current data, which is known as conditional and predictive power, and repeated confidence intervals. Similarly, the estimation of treatment difference in response could be different from the initial estimation in the design phase. This could lead to an adaptive design or sample-size re-estimation (Jennison and Turnbull, 2000).

Safety Factors: Efficacy is not the only factor that will affect a DMC's decision. Safety factors are critical for the DMC to make an appropriate recommendation to stop or continue the trial. The term "benefit-risk ratio" is a most commonly used composite criterion to assist in decision-making. In this respect, it is desirable to know the likelihood of success of the trial based on current data, i.e., the conditional power or predictive power.

The DMC may also weight the short-term and long-term treatment effects in its recommendations.

The commonly used tools for monitoring a group sequential design are stopping boundaries, conditional and predictive powers, futility index, and repeated confidence interval. These tools can be used in other adaptive designs. Bayesian monitoring tools are appreciated for different adaptive designs. There are several good books on trial monitoring: *Data Monitoring Committees in Clinical Trials* by Ellenberg, Fleming, and DeMets, 2002 and *Statistical Monitoring of Clinical Trials* by Proschan, Lan, and Wittes 2006, among others.

17.5 Analysis and Reporting

Data analyses of an adaptive design at interim analysis and at the final stage remain very challenging to statisticians. While they benefit from the

flexibility of adaptations, it is a concern that the p-value may not be correct and the corresponding confidence interval may not be reliable (EMEA, 2002). It is also a concern that major adaptations could lead to a totally different trial that is unable to address the medical questions or hypotheses that the original trial intended to answer. Although some unbiased estimators of treatment effect for some adaptive designs are available, many issues in data analysis of general adaptive designs are still very debatable from statistical and scientific points of view. It is suggested that both unadjusted and adjusted values are reported, including the adjusted and unadjusted point-estimates, the naive and adjusted confidence intervals, and the adjusted and unadjusted p-values. When an unbiased estimate is not available, computer simulations should be performed to assess the magnitude of the bias. It is suggested that in addition to the frequentist results, pharmaceutical companies should make efforts to present Bayesian results because these more-informative results could help the company itself and the FDA to better measure the benefit-risk ratio for a candidate drug.

17.6 Bayesian Approach

Bayesian methods offer a better sequential learning mechanism through Bayes' theorem, improved predictive capabilities of future results by incorporating the uncertainties of priors, and an objective evidence measure by synergizing the information resources. The Bayesian approach is information driven and decision oriented, rather than hypothesis based. Therefore, the Bayesian approach is a natural fit for adaptive designs. Bayesian approaches, in combination with adaptive methods and simulations, provide a powerful tool to achieve better planning, better design, better monitoring, and better execution of clinical trials. They can help to streamline the drug development process, increase the probability of success, reduce the cost and time-to-market for drug development, and ultimately bring the best treatment to patients faster.

17.7 Clinical Trial Simulation

Traditional drug development is subjective to a large extent, and intuitive decision-making processes are primarily based on individual experiences. Therefore, optimal design is often not achieved. Clinical trial simulation (CTS) is a powerful tool for providing an objective evaluation of development plans and study designs for program optimization and for supporting strategic decision-making. CTS is very intuitive and easy to implement

with minimal cost and can be done in a short time. The utilities of CTS include, but are not limited to (1) sensitivity analysis and risk assessment, (2) estimation of probability of success (power), (3) design evaluation and optimization, (4) cost, time, and risk reduction, (5) clinical development program evaluation and prioritization, (6) trial monitoring and interim prediction of future outcomes, (6) prediction of long-term benefit using short-term outcomes, (7) validation of trial design and statistical methods, and (8) streamlining communication among different parties. Within regulatory bodies, CTS has been frequently used for assessing the robustness of results, validating statistical methodology, and predicting long-term benefit in accelerated approvals. CTS plays an important role in adaptive design for the following reasons: First, statistical theory for adaptive designs is often complicated under some relatively strong assumptions, and CTS is useful in modeling very complicated situations with minimum assumptions not only to control type-I error, but also to calculate the power, and to generate many other important operating characteristics such as the expected sample-size, conditional power, and unbiased estimates. Second, CTS can be used to evaluate the robustness of the adaptive design against protocol deviations. Moreover, CTS can be used as a tool to monitor trials, predict outcomes, identify potential problems, and provide remedies for resolutions during the conduct of the trials.

In summary, clinical trial simulation is a useful tool for adaptive designs in clinical research. It can help investigators achieve better planning, better designs, better monitoring, and generally better execution of clinical trials. In addition, it can help to (1) streamline the drug development process, (2) increase the probability of success, and (3) reduce the cost and time-to-market in pharmaceutical research and development.

A simplified CTS model is shown in Figure 17.1. The high-level algorithms for the simulations are: (1) Simulate the trial under the null hypothesis m times. For each simulation, calculate the test statistic and use the m test statistic values to construct distribution numerically. (2) Similarly, simulate the trial under the alternative hypothesis m times, calculate the test statistic, and use the m test statistic values to construct a distribution under Ha numerically. The two distributions can be used to determine the critical region for a given α, the p-value for given data, and power for a given critical region.

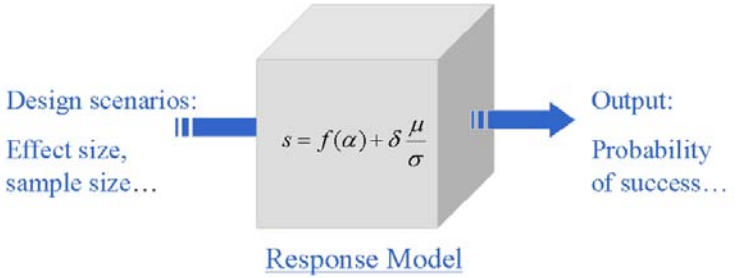


Figure 17.1: Simplified CTS Model: Gray-Box

17.8 Summary

Adaptive design methods represent new territory in drug development. Using adaptive design, we can increase the chances for the success of a trial with reduced cost. Bayesian approaches provide new tools for optimizing trial design and clinical development program planning by integrating all the relevant knowledge and information available. Clinical trial simulations offer a powerful tool to design and monitor trials. The combination of adaptive design, the Bayesian approach, and trial simulation forms an ultimate statistical instrument for most successful drug development programs.

This innovative approach requires careful upfront planning and the ability to rapidly integrate knowledge and experiences from different disciplines into the decision-making process. It requires integration of new data in real time, where data standardization tools such as CDISC and EDC play critical roles. Last but not least, all of the above require a shift to a more collaborative working environment among disciplines.

Problem

17.1 Plan an adaptive trial including all the necessary steps from beginning to the end.

Chapter 18

Paradox - Debates in Adaptive Designs

In this last chapter, we will discuss most of the controversial issues surrounding adaptive designs. We will review relevant statistical principles and check against adaptive designs. The discussions will be from both statistical and philosophical perspectives. In many ways, this may reflect the future direction.

18.1 My Standing Point

"There is only one single history, but we view it as a random sample to predict the future." — Mark Chang

"Because we cannot avoid errors, the goal is not to minimize the chance of making errors, but to minimize the impact of the errors." — Mark Chang.

“We can’t solve problems by using the same kind of thinking we used when we created them.” — Albert Einstein

"We insist on the fact that statistics should be considered an interpretation of natural phenomena, rather than an explanation.” — Christian Robert.

There are different paradigms of statistical theory. The different theories represent different philosophies and have provoked much controversy. However, within each paradigm, consistency and completeness are expected with an axiom system. These axioms are often called principles. When a new inference procedure or experiment design violates a principle of an existing theory, there are several possible actions we can take: (1) Do not use the new procedure; (2) Modify the existing theory or abandon the "principle" so we can explain a new "phenomena" or use a new procedure; or (3) Develop a new theory or choose another existing theory.

We cannot completely avoid making errors. Taking clinical trials as an example, we can make a type-I (false positive) or type-II (false negative) error. If one believes that protecting patients from ineffective drugs is of absolute importance, one should use $\alpha = 0$ for the hypothesis test so that no drug (effective or ineffective) will get to the market. On the other hand, if one believes that preventing any effective drug from being delivered to patients is absolutely unacceptable, one should use $\alpha = 1$ so that no drug (effective or ineffective) will be left out. In reality, we should balance the effects of making these two types of errors. In other words, what we really care about is the impact of errors, not the errors themselves.

In statistics, there are many principles; any violation of these principles can be considered an "error" and will have an impact. We all know that in mathematics, statistics, and science, there are many theories that conflict with one another, even though each is internally consistent. In statistics, for example, Bayesian and frequentist approaches have many fundamental differences. An adaptive design may violate principles of one theory, but may be consistent with the principles of other theories. Our action taken, e.g., choosing either classic or adaptive design, is dependent on the impact of that decision.

Statistical analyses and predictions are usually motivated by objectives. The outcome of the analyses and predictions will guide the decision-making. When we propose a statistical method (e.g., adaptive design) under a certain condition, we believe the method is preferable to alternative methods based on anticipated consequences or impacts.

The impact is characterized by a loss function in decision theory. This loss function (implicit or explicit) always guides our decision whether we realize it or not. The challenging and also interesting part is that different people have different perspectives on loss, hence the different loss functions. Loss functions are often vague and not explicitly defined, especially when we make decisions in our daily lives. Decision theory make this loss explicit and deals with it with mathematical rigor.

18.2 Decision Theory Basics

In decision theory, statistical models involve three spaces: the observation space X, the parameter space Θ, the action space A. Actions are guided by a decision rule $\delta(\mathbf{x})$. An action $\alpha \in A$ always has an associated consequence characterized by the so-called loss function $L(\theta, a)$. In hypothesis testing, the action space is $A = \{\text{accept; reject}\}$.

Because it is usually impossible to uniformly minimize the loss $L(\theta, a)$, in frequentist paradigm, the decision rule δ is determined such that minimize the following average loss:

$$\begin{aligned} R(\theta, \delta) &= E^{X|\theta}(L(\theta, \delta(\mathbf{x}))) \\ &= \int_X L(\theta, \delta(\mathbf{x})) f(\mathbf{x}|\theta) d\mathbf{x}. \end{aligned} \tag{18.1}$$

The rule $a = \delta(\mathbf{x})$ is often called an estimator in estimation problems. Common examples of loss function are *squared error loss* (SEL) $L(\theta, a) = (\theta - a)^2$, *absolute loss*, $L(\theta, a) = |\theta - \alpha|$, the *0-1 loss*, $L(\theta, a) = \mathbf{1}(|\alpha - \theta|)$, etc.

The expected SEL associate variance and bias of an estimator through following:

$$E^{X|\theta}(\theta - \delta(X))^2 = Var(\delta(X)) + [\text{bias}(\delta(X))]^2,$$

where $\text{bias}(\delta(X)) = E^{X|\theta}(\delta(X))$.

Definition 18.1 Bayesian expected loss is the expectation of the loss function with respect to posterior measure, i.e,

$$\rho(\delta(\mathbf{x}), \pi) = E^{\theta|X} L(\delta(\mathbf{x}), \theta) = \int_\Theta L(\theta, \delta(\mathbf{x})) \pi(\theta|x) d\theta \tag{18.2}$$

An action $a^* = \delta^*(\mathbf{x})$ that minimizes the posterior expected loss is called Bayes action.

By averaging (18.1) over a range of θ for a given prior $\pi(\theta)$, we can obtain:

$$\begin{aligned} r(\pi, \delta) &= E^{\pi}(R(\theta, \delta)) \\ &= \int_\Theta \int_X L(\theta, \delta(\mathbf{x})) f(\mathbf{x}|\theta) \pi(\theta) d\mathbf{x} d\theta. \end{aligned} \tag{18.3}$$

The two notions by (18.2) and (18.3) are equivalent in the sense that they lead to the same decision.

Theorem 18.1 An estimator minimizing the integrated risk $r(\pi, \delta)$ can be obtained by selecting, for every $x \in X$, the value $\delta(x)$ which minimizes the posterior expected loss, $\rho(a, \pi)$, because

$$r(\pi, \delta) = \int_X \rho(a, \delta(\mathbf{x})|\mathbf{x}) m(\mathbf{x}) d\mathbf{x}.$$

18.3 Evidence Measure

Before we can assess the impact of an error, we have to understand it profoundly. Remember, an error can be viewed as a negation of evidence. Furthermore, drug development is a sequence of decision-making processes, and decisions are made on the basis of evidence. Therefore, the way in which evidence is measured is critical. The totality of evidence indicating that a drug is beneficial to the patient population is a very complex issue and involves many aspects. However, for simplicity, we will focus on efficacy evidence. We are going to discuss four different measures of evidence: p-value, likelihood, Bayes' factor, and the Bayesian p-value.

18.3.1 *Frequentist P-Value*

The p-value defined by the conditional probability Pr(data|Ho) is the error rate of claiming efficacy, when in fact there is no treatment effect (i.e., the null hypothesis Ho is true). In drug development, the ultimate question is the following: What is the treatment effect given the observed data? That is the exact question that the Bayesian method answers using posterior probability Pr(Ha|data). However, the frequentist p-value for hypothesis testing only tells us the probability of observing at least this treatment difference given that the null hypothesis is true. In other words, p-value is a measure of evidence against the null. In this sense, the frequentist approach doesn't answer the question.

There have been extensive discussions on p-value (Fisher, 1999). For example, it is believed that there is always a difference between any two treatments (though it might be very small). Therefore, as long as the sample-size is large enough, statistical significance will be demonstrated. The p-value depends on the power. Furthermore, p-value does not provide a consistent measure of evidence because identical p-values do not imply identical evidence of treatment effect. Suppose two trials have been conducted. One trial shows a mean treatment difference of 5 with a CI (0, 10) and p-value of 0.05, and the other shows a mean treatment difference of 50 with a CI (0,100) and p-value of 0.05. Clearly, the second trial has provided stronger evidence supporting the effectiveness of the test drug, even though the p-values from the two trials are identical.

18.3.2 *Maximum Likelihood Estimate*

Likelihood is an important concept in both frequentist and Bayesian paradigms, and can be illustrated as follows: Given the observed data y in the

current trial, the joint probability distribution function $p(y, \theta)$ can be considered a function of the treatment effect θ (e.g., median survival difference between two groups), and tells us how strongly the data support different θs. When $p(y, \theta)$ is viewed this way, it is known as the likelihood function. The likelihood function measures the relative plausibility of different values of θ. The value of θ corresponding to the maximum value of the likelihood function is called the maximum likelihood estimate (MLE) of θ.

Suppose in a single-group oncology trial, 12 responses were observed out of 30 patients. The likelihood function can be obtained based on the observed data (Figure 18.1). The maximum likelihood estimate for response rate is 40%. The MLE is consistent with the naive estimate, i.e., $12/30 = 40\%$.

Maximum likelihood is commonly used to estimate the median survival time when there are censored data. However, MLE is not a good single measure of evidence because it does not take the other possible median survival times into consideration.

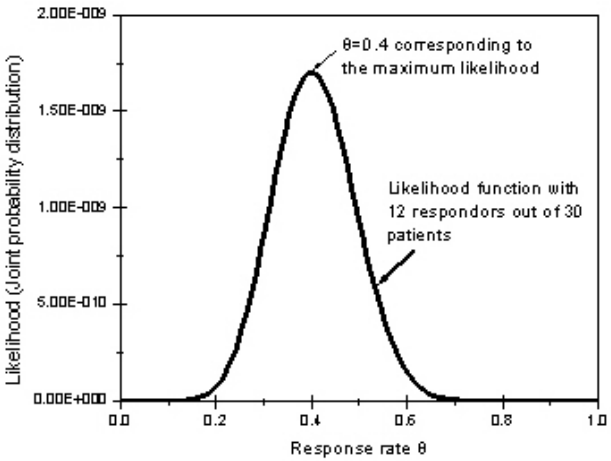


Figure 18.1: Illustration of Likelihood Function

18.3.3 *Bayes Factor*

As discussed earlier, p-value and MLE alone do not provide a good measure for evidence. For a hypothesis test of the null hypothesis H_o (no treatment effect θ) versus the alternative hypothesis H_a (with treatment effect θ), if

we consider both the evidence supporting the null hypothesis being true and the evidence supporting the alternative hypothesis being true, an intuitive way to compare them is to take the ratio of the two evidences, the so-called Bayes' factor (BF).

Definition 18.2 The Bayes factor is the ratio of the posterior probabilities of the null and alternative hypotheses over the ratio of the prior probabilities of the null and the alternative hypotheses, i.e.,

$$BF(x) = \frac{P(\theta \in \Theta_o|x)}{P(\theta \in \Theta_a|x)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_a)}, \tag{18.4}$$

or

$$BF(x) = \frac{\int_{\Theta_o} f(x|\theta_o)\,\pi_o(\theta)\,d\theta}{\int_{\Theta_a} f(x|\theta_a)\,\pi_a(\theta)\,d\theta}. \tag{18.5}$$

From (18.4) we know when $\pi(\theta \in \Theta_0) = \pi(\theta \in \Theta_a)$, BF = likelihood ratio.

$$BF = (\text{Likelihood for } \theta \text{ in } Ho)/(\text{Likelihood for } \theta \text{ in } Ha)$$

Reject H_o if $BF \leq k$, where k is a small value, e.g., 0.1. A small value of BF implies strong evidence in favor of H_a or against H_o. Note that there are several slightly different definitions of Bayes' factor.

Because of the different methods of measuring the evidence, the same data can lead to different conclusions, which has been stated in Lindley's paradox: When the information ratio (information from current trial versus prior information) is high and p-value is just marginally significant against H_o, the BF can be greater than 1, and hence support H_o (Lindley, 1957; Spiegelhalter et al., 2004)

18.3.4 *Bayesian P-Value*

Suppose we want to test the null hypothesis that $H_o : \theta = 0$ against an alternative hypothesis that $H_a : \theta > 0$.

The frequentist p-value is the probability of having the observed treatment difference or larger assuming $\theta = 0$, i.e., $Pr(data \text{ or more extreme}|\theta = 0)$. The Bayesian p-value, calculated from posterior distribution is the probability of $\theta \leq 0$ (no treatment effect) given the observed data, i.e., $Pr(\theta \leq 0|data)$.

Isn't the Bayesian p-value what we are desperately looking for, and not the frequentist p-value? Bayesian significance is claimed when the Bayesian p-value is less than or equal to the so-called Bayesian significance level α, a predetermined constant that does not have to be the same as the frequentist significance level.

18.3.5 *Repeated Looks*

One of the difficulties in using the frequentist approach is the issue of multiple testing, or repeated looks. When multiple tests are performed, such as in trials with multiple endpoints and in adaptive trials or pharmacogenomic studies, the false positive rate will be inflated when using naive methods. Therefore, p-value adjustment or alpha penalty should be applied in such cases. The problem is that as the number of analyses gets larger, the power for detecting a difference diminishes quickly due to the multiplicity adjustment.

In fact, when the number of looks approaches infinity, the probability of making type-I error is 100%. However, the probability of the BF being less than α is always less than α, regardless of the number of looks, i.e., $Pr(BF < \alpha | H_o) \leq \alpha$. This theorem states that when the null hypothesis is true, if we look repeatedly for a BF of less than 10%, strong evidence against the null, this will not occur more than 10% of the time, no matter how many times we look (Royall, 1997). This theorem thus reveals the maximum probability of misleading evidence, but only if we measure the evidence properly.

Because of the controversial issues surrounding multiplicity in classic hypothesis testing based on the Neyman-Pearson theory, the frequentist p-value has been strongly criticized from the Bayesian and Fisherian perspectives (Spiegelhalter, et al., 2004). Cornfield questioned ironically: "Do we want error control over a single trial, over all the independent trials on the same agent, on the same disease, over the lifetime of an investigator, etc.?" (Cornfield, 1976). Each of these control methods can lead to different adjusted p-values and hence contradictory statistical conclusions.

18.3.6 *Role of Alpha in Drug Development*

As we all know, the frequentist is fundamentally based on (infinitely) repeated experiments. However, we virtually never repeat (many times) the pivotal clinical trial for the same compound for the same indication. The real replications are the clinical trials of different compounds for the same or different indications. Therefore, we should take a close look at the im-

plications caused by a fixed α enforced by regulatory body.

From a regulatory point of view, the statistical criterion of a p-value $<\alpha$ can be viewed as a simple tool to control the proportion of ineffective drugs on the market. However, that is just one side of the story. We also have to consider the downside, which is preventing good drugs from being delivered to patients because of this criterion. To balance the two sides, we can use a benefit-risk ratio or utility. Bayesian decision theory is a powerful tool in this regard. Furthermore, using a fixed α of 5% does not necessarily control the release of ineffective drugs onto the market. For example, if all drug candidates in phase III are efficacious, all drugs on the market will be effective regardless of α. Similarly, if all drug candidates in phase III are inefficacious, then all drugs on the market will be ineffective regardless of α. The problem is that we don't know the proportion of effective drug candidates in phase III trials unless the Bayesian approach is adopted.

The frequentist paradigm has played an important historical role in drug development. The frequentist approach, with alpha control at the 5% level, was appropriate because there were many compounds that had major effects. This high standard (error rate <5%) for drug approval has led to the most effective compounds being selected and approved, and at the same time it has prevented a larger number of ineffective drugs from spreading into the market. The frequentist criterion was probably consistent with the benefit-risk measure, had it been developed. However, the situation has changed: given that there are so many drugs already on the market, the margin for improvement is getting smaller, and an active controlled study requires an extremely large sample-size that is often not feasible. Even if there are a few remaining drugs with large effects, they will be hard to find using the traditional frequentist method. Given these reasons, the statistical criterion of alpha = 5% could require an unreasonably large benefit-risk ratio. The Bayesian approach is a better alternative, despite its challenges; after all, challenges are the force that drives science forward. Personalized medicine is the future for the patients, however, to be able to effectively develop personalized medicine, advancement in science and technology for drug development is critical.

18.4 Statistical Principles

A main purpose of statistical theory is to derive an inference about the probability distribution from observations of a random phenomenon. The distribution model is simple and often an efficient way to describe a past phenomenon and more importantly to predict a future event of a similar

nature. David Cox (2006) gives an excellent review of important statistical principles from both frequentist and Bayesian perspectives in his recent book: *Principles of Statistical Inference.*

Definition 18.3 A parametric statistical model consists of the observation of a random variable x, distributed according to $f(x|\theta)$, where only the parameter θ is unknown and belongs to a vector space Θ of finite dimension.

Once the statistical model is defined, a main task of the statistical analysis is to lead to an inference (estimation or hypothesis testing) on the parameter θ. In contrast to the probabilistic modeling, the purpose of a statistical analysis is fundamentally an inversion purpose, since it aims at retrieving the "causes", i.e., parameters of the probabilistic generating mechanism, from the "effects" or observations. In early time, statistics is also called "Inverse Probability."

The fiducial approach of Fisher (1956) also relies on this inversion. Let's denote C the cause and E the effect. Considering the relation $E = C + \varepsilon$ where ε is an error term. It is argued that, if C is known, E is distributed according to the above relation. Conversely, if E is known, $C = E - \varepsilon$ is distributed according to the symmetric distribution. However if E is a random variable and C is a (constant) parameter, to write $C = E - \varepsilon$ does not make that C a random variable. The fiducial approach was abandoned after the exposure of fundamental paradoxes (Stein, 1959; Robert, 1997).

Definition 18.4 A Bayesian statistical model is made of a parametric statistical model, $f(x|\theta)$, and a prior distribution on the parameters, $\pi(\theta)$.

Theorem 18.2 (**Bayes' Theorem)** Bayes theorem can be expressed as

$$\pi(\theta|x) = \frac{f(x|\theta)\,\pi(\theta)}{\int f(x|\theta)\,\pi(\theta)\,d\theta}. \tag{18.6}$$

Bayes's Theorem places causes (observations) and effects (parameters) on the same conceptual level, since both of them have probability distributions. It is considered as a major step from the notion of an unknown parameter to the notion of a random parameter (Robert, 1997). However, it is important to distinguish x and θ, x is observable, but θ is latent.

Definition 18.5 When $x \sim f(x|\theta)$, a function T of x (also called a statistic) is said to be *sufficient* if the distribution of x conditionally on $T(x)$ does not depend on θ.

A sufficient statistic $T(x)$ contains the whole information brought by x about θ.

Theorem 18.3 (Neymam) (Hogg, McKean, and Graig, p.376). Let $X_1, ..., X_n$ denote a random sample from a distribution that has p.d.f. or p.m.f. $f(x;\theta)$, $\theta \in \Omega$. The statistic $Y_1 = T_1(X_1, ...X_n)$ is a sufficient statistic for θ, if and only if there exist two nonnegative functions, η_1 and η_2, such that

$$f(x_1;\theta)...f(x_n;\theta) = \eta_1[T_1(x_1, ..., x_n);\theta]\,\eta_2(x_1, ..., x_n), \tag{18.7}$$

where $\eta_2(x_1, ..., x_n)$ does not depend upon θ.

When the model allows for a minimal sufficient statistic, i.e., for a sufficient statistic which is a function of all the other sufficient statistics, we only have to consider the procedures depending on this statistic.

Sufficiency Principle Two observations x and y factorizing through the same value of a sufficient statistic T, i.e., such that $T(x) = T(y)$, must lead to the same inference on θ.

Christian Robert (1997) pointed out: "The Sufficiency Principle is only legitimate when the statistical model is actually the one underlying the generation of the observations. Any uncertainty about the distribution of the observations should be incorporated into the model, a modification which almost certainly leads to a change of sufficient statistics. A similar cautionary remark applies to the Likelihood Principle."

Conditionality Principle If m experiments $\left(\breve{E}_1, ..., \breve{E}_m,\right)$ on the parameter θ are available with equal probability to be selected, the resulting inference on θ should only depend on the selected experiment.

Stopping Rule Principle If a sequence of experiments, $\breve{E}_1$, $\breve{E}_2, ...$, is directed by a stopping rule, τ, which indicates when the experiments should stop, inference about θ must depend on τ only through the resulting sample.

The Bayesian decision is independent of the stopping criterion, therefore is not influenced by the subjective motivations which led to the resulting sample-size. For example, in a clinical trial, from sponsor perspective, the trial can continue to recruit patients until p-value$<\alpha$. The problem is that the regulatory body, physicians and patients, have different loss functions. If the regulatory body has the final say about the loss function, then we unlikely have the stopping rule that allows a trial to continue recruiting

patients until p-value $< \alpha$. In such case, p-value $< \alpha$ can be viewed as a constraint in the minimization using decision theory.

Likelihood Principle The information contained by an observation x about θ is entirely contained in the likelihood function $l(\theta|x)$. Moreover, if x_1 and x_2 are two observations depending on the same parameter θ, such that there exists a constant c satisfying

$$l_1(\theta|x) = cl_2(\theta|x_2) \tag{18.8}$$

for every θ, they contain the same information about θ and must lead to identical inferences.

Note that the Likelihood Principle is only valid when (i) inference is about the same parameter θ and (ii) θ includes every unknown factor of the model.

Likelihood principle can be challenged, e.g., why should the likelihood principle be in the multiplicity form of (18.8), instead of additivity form:

$$l_1(\theta|x) = l_2(\theta|x_2) + c.$$

Example 18.1 Paradox: Binomial and Negative Binomial?

Suppose, we are interested in the hypothesis testing of a binary endpoint.

$$H_o : p = 0.5 \text{ vs: } H_a : p > 0.5.$$

The experiment is finished with 3 responses out of 12 patients. However this information is not sufficient for rejecting or accepting the null hypothesis.

Scenario 1: If the total number of patients, N = 12 is predetermined, the number of responses X follows binomial distribution $B(n; p)$, the frequentist p-value of the test is given by

$$\Pr(X \geq 9|H_o) = \sum_{x=9}^{12} \binom{12}{x} 0.5^x 0.5^{12-x} = 0.073.$$

The null cannot be rejected at a one–sided level $\alpha = 0.05$. The likelihood in this case is given by

$$l_1(x|p) = \binom{12}{9} p^9 (1-p)^3 = 220p^9(1-p)^3.$$

Scenario 2: If the number of response, $n = 3$, is predetermined and the experiment continues until 3 responses are observed, then X follows negative binomial $NB(3; 1-p)$ and the frequentist p-value of the test is given by

$$\Pr\left(X \geq 9|H_o\right) = \sum_{x=9}^{\infty} \binom{3+x-1}{2} 0.5^x 0.5^3 = 0.0327,$$

because $\sum_{x=k}^{\infty} \binom{2+x}{2} \left(\frac{1}{2}\right)^x = \frac{8+5m+m^2}{2^m}$. Therefore the null is rejected at a one-sided level $\alpha = 0.05$. The likelihood in this case is given by

$$l_2\left(x|p\right) = \binom{3+9-1}{2} p^9 \left(1-p\right)^3 = 55p^9\left(1-p\right)^3.$$

According to Likelihood Principle, all relevant information is in the likelihood $l\left(p\right) = p^9\left(1-p\right)^3$ and therefore the two scenarios should not lead to different conclusions (rejection or not rejection).

Theorem 18.4 The Likelihood Principle is equivalent to the conjunction of the Sufficiency and the Conditionality Principles.

Proof (Robert, 1997, p.18). We prove the case for $m = 2$. Let's denote the evidence associated with an experiment $\breve{E}$ by $Ev\left(\breve{E}, x\right)$, as the collection of the possible inferences on the parameter θ directing this experiment. Let $\breve{E}^*$ denote the mixed experiment starting with the choice of $\breve{E}$ with probability 0.5 ($i = 1, 2$), thus with result (i, x_i). Under these notations, the Conditionality Principle can be written as

$$Ev(\breve{E}^*, (j, x_j)) = Ev(\breve{E}_j, x_j). \tag{18.9}$$

Consider x_1^0 and x_2^0 with likelihood functions such that

$$l_1\left(\cdot|x_1^0\right) = cl_2\left(\cdot|x_2^0\right). \tag{18.10}$$

The Likelihood Principle then implies

$$Ev(\breve{E}_1, x_1^0) = Ev(\breve{E}_2, x_2^0). \tag{18.11}$$

Let's assume that (18.10) is satisfied. For the mixed experiment $\breve{E}^*$ derived from the two initial experiments, consider the statistic

$$T(j, x_j) = \begin{cases} (1, x_1^0) & \text{if } j = 2,\ x_2 = x_2^0 \\ (j, x_j) & \text{otherwise} \end{cases},$$

which takes the same value for $(1, x_1^0)$ and for $(2, x_2^0)$. Then, this statistic is sufficient since, if $t \neq (1, x_1^0)$,

$$P_\theta\left(X^* = (j, x_j) \,|T = t\right) = \mathrm{II}_t\left(j, x_j\right)$$

and

$$P_\theta\left(X^* = \left(1, x_1^0\right) |T = \left(1, x_1^0\right)\right) = \frac{c}{1+c},$$

due to the proportionality of the likelihood functions. The Sufficiency Principle then implies that

$$Ev(\breve{E}^*, (1, x_1)) = Ev(\breve{E}^*, (2, x_2)), \tag{18.12}$$

and, combined with (18.10), leads to (18.11).

The reciprocal of this theorem can be derived for the Conditionality Principle from the fact that the likelihood functions of (j, x_j) and x_j are proportional and for the Sufficiency Principle from the factorization theorem.

The Likelihood Principle does not lead any operational procedure for hypothesis testing. There are several ways to make it operational. For example, maximum likelihood estimation method is an operational enhancement of the Likelihood Principle. The Baye's theorem can be viewed as an operational enhancement of the Likelihood Principle.

Maximum Likelihood Estimator (MLE) When $x \sim f(x|\theta)$ is observed, the maximum likelihood estimator of θ is defined as:

$$\hat{\theta} = Arg\max l(\theta|x), \tag{18.13}$$

where the notation $Arg\max$ means that the likelihood $l(\theta|x)$ achieves its maximum value at $\hat{\theta}$.

Note that the maximization (18.13) can lead to several maxima. The maximum likelihood estimator is found useful because (1) it's intuitive motivation of maximizing the probability of occurrence; (2) it has strong as-

ymptotic properties (consistency and efficiency); (3) it is parametrization-invariant.

Invariance Principle If $\hat{\theta}$ is the maximum likelihood estimator, then for any function $h\left(\hat{\theta}\right)$, the maximum likelihood estimator of $h(\theta)$ is $h(\hat{\theta})$ (even when h is not one-to-one).

This property is not enjoyed by most other statistical approaches. However, maximum likelihood estimates are often biased. Does it imply that the world is operated biasely?

MAP estimator MAP (maximum aposteriori) estimator is also called the generalized maximum likelihood estimator (GMLE). The GMLE is the largest mode of the $\pi(\theta|x)$. The MLE maximizes $l(\theta)$, while the GLME maximizes $\pi(\theta)\, l(\theta)$.

18.5 Behaviors of Statistical Principles in Adaptive Designs

18.5.1 *Sufficiency Principle*

Jennison and Turnbull (2003), Posch et al. (2003), and Burman and Sonesson (2006) pointed out that the adaptive design using a weighted statistic violates the sufficiency principle, because it weights the same amount of information from different stages differently. Michael A. Proschan (Burman and Sonesson, 2006, discussion) proved that no test based on the sufficient statistic can maintain level α irrespective of whether the prespecified sample-size rule is followed. Thach and Fisher (2002, p. 436) and Burman and Sonesson (2006) highlighted the problem of very different weights. An example (given by Marianne Frisen, in Burman and Sonesson, 2006, discussion) of a violation of the sufficiency principle is the use of the median instead of the mean when estimating the expected value of a normal distribution.

The "unweighted test" utilizes the distribution of N to choose a critical level with the desired probability of rejecting the null hypothesis unconditional on N but disregards the observed value of N. Because N is part of the minimal sufficient statistic, it is not in accordance with the sufficiency principle and is inefficient (Marianne Frisen, in Burman and Sonesson, 2006, discussion).

Should information (or sufficiency of a statistic) regarding θ be based on data or data + procedure? We probably can argue that the sufficiency of

a statistic should not only depend on the data, but also on the procedure or experiment used to collect the data. The data can contain different amounts of information if they are collected differently. The conclusion that each observation contains the same amount of information is valid in a classic design, but not in an adaptive design, because a later observation contains some information from previous observations due to the dependent sampling procedure. Therefore, an insufficient statistic in classic design can become sufficient under an adaptive design. For example, suppose we want to assess the quantity $\vartheta = \frac{1-R_B}{1-R_A}$, where R_B and R_A are the true response rates in groups A and B, respectively. In a classic design with balanced randomization $\frac{N_A}{N_B}$ is asymptotically approaching 1. However, in the randomized-play-the-winner design, $\frac{N_A}{N_B}$ is asymptotically equal to ϑ. Therefore the same data $\frac{N_A}{N_B}$ could contain different amounts of information in different designs.

It should be apparent that a statistic (e.g., $\bar{X}$) can be a sufficient statistic conditionally but not unconditionally for θ.

18.5.2 *Minimum Sufficiency Principle and Efficiency*

Jennison and Turnbull (2006) raised the question as to when an adaptive design using nonsufficient statistics can be improved upon by a nonadaptive group sequential design. Tsiatis and Mehta (2003) have proved that for any SSR adaptive design, there exists a more powerful group sequential design; however, the group sequential test has to allow analyses at every cumulative information level that might arise in the adaptive design. On the other hand, the weighted method (or MINP), provides great flexibility and when the sample-size deos not change, it has the same power as the classic group sequential design.

It is really analogous with a two-player game. Player A says if you tell me the sample-size rule, I can find a group sequential design (with as many analyses as I want to have) that is more efficient than the adaptive design. Player B says, if you tell me how many and when you plan the interim analyses, I can produce an adaptive trial with the same number and timing of the analyses that is more flexible, and in case the sample-size does not change, it is identical to the group sequential design.

Interestingly, a group sequential design can also be viewed as a special SSR adaptive design. The adjustment rule at interim analysis is that sample-size will not be increased if futility or efficacy is found; otherwise add n_2 subjects. Therefore, what Tsiatis and Mehta eventually proved is that there is an SSR design with a discrete sample-size increment that is more efficient than a given SSR with continual sample-size increment.

Uniformly a more powerful design has an important implication: If design A is uniformly more efficient or powerful than design B, then $\{\delta; Ho$ is not rejected with design $A\} \subset \{\delta; Ho$ is not rejected with design $B\}$. Because the confidence interval consists of all values of δ for which H_o is not rejected, $CI_A \subseteq CI_B$.

Jennison and Turnbull (2006) stated, "A design is said to be inadmissible if another design has a lower expected information function and higher power curve. A design that is not inadmissible is said to be admissible. The fundamental theoretical result is a complete class theorem, which states that all admissible designs are solutions of Bayes sequential decision problems. Because Bayes designs are functions of sufficient statistics, any adaptive design defined through nonsufficient statistics is inadmissible and is dominated by a design based on sufficient statistics. This conclusion confirms that violation of the sufficiency principle has a negative impact on the efficiency of an adaptive design." The question is should the information and consequently the sufficiency be procedure dependent?

18.5.3 *Conditionality and Exchangeability Principles*

Burman and Sonesson (2006) pointed out that SSR also violates the invariance and conditionality principles because the weighted test depends on the order of exchangeable observations. Marianne Frisen (Burman and Sonesson, 2006, discussion) stated similarly: "The 'weighted' test avoids this by forcing a certain error spending. This is done at the cost of violating the conditionality principle. The ordering of the observations is an ancillary statistic for a conclusion about the hypothesis. Thus, by the conditionality principle the test statistic should not depend on the ordering of the realized observations." For a general distribution, not belonging to the exponential family, the weighted test will violate the conditionality principle but not the sufficiency principle (Burman and Sonesson, 2006).

If sufficiency and conditionality are important, then the combination of the two is the likelihood principle, which also faces challenges (see examples provided earlier in this chapter). The bottom line is that the world is unpredictable by humans, but it is also deterministic. When it is viewed as random, there will be many paradoxes.

When we talk about exchangeability, we should apply the same data scope to both classic and adaptive designs. The same terminology, "experiment-wise," can imply a different data scope for classic and adaptive designs. For example, in a phase II and III combined seamless adaptive design, data from phases II (or the learning phase) and III (the confirmatory phase) are combined and viewed as experiment-wise data, while in a classic

design, the data from the two phases are analyzed separately simply because we call them two different experiments. Fairly, if we compare the two design approaches at the same data scope, i.e., phase II and III combined, then we will see a very interesting result: the classic design is a special case of an adaptive design using the weighted test, where the weight for the "learning phase" is zero and the weight for the "confirmatory phase" is 1.

$$\text{Classic design} = 0\ \text{(learning-phase data)} + 1\ \text{(confirmatory-phase data)}$$

$$\text{Adaptives design} = w_1\ \text{((learning-phase data)} + w_2\ \text{(confirmatory-phase data)}$$

Humans are not memory-less creatures; we should not pretend that we don't have the learning-phase data when we analyze the confirmatory-phase data. Data from different phases of an adaptive trial is often not exchangeable; data from two trials of different phases are not exchangeable in a classic design. Is it just a terminological difference: "phase" or "trial"?

18.5.4 *Equal Weight Principle*

Exchangeable observations should be weighted equally in the test statistic. The equal weight principle views exchangeability from the weight perspective. However, the measurement of the information level is not unique. For example observation x_i can be counted as a unit of information, $1/x_i$ can be counted as a unit of information, or p_i can be counted as unit of information. The adaptive design method MSP equally weights the evidences (p_i) against H_o from the two stages.

From an ethical point of view, should there be equal weight for everyone, one vote for one person? Should efficacy be measured by the reduction in number of deaths or by survival time gained? Should it be measured by mean change or percent change from baseline? All these scenarios apply a different "equal weight" system to the sample. Suppose you have a small amount of a magic drug, enough to save only one person in a dying family: the grandfather, the young man, or the little boy. What should you do? If you believe life is equally important for everyone regardless of age, you may randomly (with equal probability) select a person from the family and give his/her the drug. If you believe the amount of survival time saved is important (i.e., one year of prolonged survival is equally important to everyone), then you may give the drug to the little boy because his life expectancy would be the longest among the three family members. If you believe that the impact of a death on society is most important, then

you may want to save the young man, because his death would probably have the most negative impact on society. If these different philosophies or beliefs are applied to clinical trials, they will lead to different endpoints and different test statistics with different weights.

Exchangeability may be only an approximation and never exist in reality. In the real word, humans take in information in an order-dependent way, because this is a better method of gaining knowledge. As a simple example, you would learn addition before multiplication. In other words, an order-dependent measure is a better measure of information level than an order-independent measure. The same knowledge, e.g, $1 + 2 = 3$, has different a information level for a preschooler and a middle school student. If the trial data are collected sequentially, why should we insist the exchangeability and pretend the data were collected instantly? The statement $1 + 2 \neq 2 + 1$ is true in many ways. For example "$1 + 2$" could mean a 200% increment in our knowledge, while "$2 + 1$" may mean a 50% increment.

18.5.5 *Consistency of Trial Results*

Compared to classic design, adaptive designs naturally allow for an interim look to check the consistency of results from different stages. If p_1 and p_2 are very different, we may have to look at the reasons, such as baseline difference, gene difference, etc. The question is should we check this consistency for a classic design too by splitting the data in different ways?

It is also controversial in adaptive designs (including group sequential designs) that we often reject the null hypothesis with less-strong evidence, but don't reject the null with stronger evidence. For example, in a two-stage GSD with the O'Brien-Fleming spending function ($\alpha_1 = 0.0025$ and $\alpha_2 = 0.0238$), we will not reject the null when the p-value $= 0.003 > \alpha_1$ at the interim look, but we do reject the null when the p-value $= 0.022 < \alpha_2$. Why don't we reject the null hypothesis when the evidence against the null is stronger (p-value $= 0.003$) and reject H_o when the evidence is much weaker (p-value $= 0.022$)?

Unless the test statistic is a monotonic function of the parameter estimator regardless of how the data were collected, it is always possible to have conflicting results. For example, median and mean have conflicting results, i.e., a rank test statistic conflicts with that from a parametric approach.

18.5.6 *Bayesian Aspects*

Two fundamental principles are naturally followed by the Bayesian paradigm with no constraint on the procedures to be considered, namely the Likelihood Principle and the Sufficiency Principle. On the other hand, the Bayesian approach rejects other principles, like the notion of unbiasedness. This notion was once a cornerstone of classical statistics and restricted the choice of estimators to those that are, on average, correct (Lehmann, 1983). Our objective is to minimize the impact of errors not the number of errors.

From the frequentist perspective, the most convincing argument in favor of the Bayesian approach is that it intersects widely with the three notions of classical optimality, namely, minimaxity, admissibility, and equivariance. Most estimators that are optimal according to one of these criteria are Bayes estimators or limits of Bayes estimators (the notion of limit depends on the context). Thus, not only is it possible to produce Bayes estimators that satisfy one, two, or three of the optimality criteria, but more importantly, the Bayes estimators are essentially the only ones that achieve this aim (Robert, 1997, Chapters 1 and 10).

The Bayes estimators use uniform representations under loss functions, while the maximum likelihood method does not necessarily lead to an estimator. For example, this is the case for normal mixtures, where the likelihood is not bounded.

An interesting question for a Bayesian approach is, should that which we learn from incremental information many times be the same as that which we learn from the cumulative information all at once? The answer is that it is dependent on the model — the answer is yes for conjugate models, but not for general models.

18.5.7 *Type-I Error, P-value, Estimation*

Type-I error control is not very challenging, and most adaptive designs control the family-wise error. However, because hypothesis testing is primarily based on the concept of repeated experiments, in what scope the experiment will potentially be repeated is critical. In clinical trials, we virtually never test the same compound for the same indication in a phase III study repeatedly for many times. For this reason, the implication of control of the experimental error rate α may be totally different from what we initially intended.

The definition, not the calculation, of p-value is challenging, because there are so many options for an adaptive design. Remember that p-value is the probability of the test statistic under the null hypothesis being more

extreme than the critical point. The key lies in how the extreme should be defined. Unlike in a classic design, definitions of extremeness of a test statistic can be in many different ways in adaptive designs, e.g., stagewise-ordering, sample mean ordering, and likelihood ratio ordering (see Exercise 8.1). None of these definitions satisfies simultaneously the concerns raised from statistical, scientific, and ethical angles.

Another relevant question is: should the duality principle between hypothesis testing and the confidence interval be applied to adaptive designs, i.e., the confidence interval consists of all δ_0 that do not cause rejection of the null hypothesis? If yes, there are many different definitions of confidence intervals, just like the p-values.

Regarding the issues in overall direction of estimation, a very narrow continual band may cause issues. Example: for a group sequential design with stopping boundaries $\alpha_1 = 0$, $\beta_1 = 0.025$, $\alpha_2 = 1$, any data from stage 2 will lead to rejection of Ho regardless of the overall direction of the treatment effect. There are also many other examples of contradictory results between weighted and unweighted methods (one positive and the other negative).

Regarding unbiased estimation, we have proven that the fixed weight method will lead to an unbiased point estimate if the trial does not allow for early stopping. If the trial allows for early stopping, the estimate is biased. What if the trial is designed for no early stopping, but is actually stopped for futility. In such a case, we don't care because the test compound wouldn't be marketed. This implies that we can always design a trial with no stopping and if it continues, we can calculate the naive estimate and claim that it is unbiased.

These controversies encourage us to examine the classic principles more carefully and adapt to the new phenomenon.

18.5.8 *The 0-2-4 Paradox*

Let's study further the type-I error, point estimate, confidence interval, and p-value through an interesting paradox, called the 0-2-4 paradox.

The 0-2-4 paradox: An experiment is to be conducted to prove that spring water is effective compared to a placebo in certain disease population. To carry out the trial, I need a coin and up to 4 patients. To control type-I error, the coin is used, i.e., **0** patient is needed; To have a unbiased point estimate, a sample-size of **2** patients is required; To obtain the confidence interval and p-value, **4** patients are sufficient. The trial is carried out as follows:

(1) To control the type-I error at one-sided $\alpha = 0.05$, I flip the coin

100 times. If the number of heads $n(\text{head}) < 95$, the trial will be stopped without efficacy claim. For this, I don't even need either the water or patients. If the number of heads $n(\text{head}) \geq 95$, the efficacy will be claimed and the trial proceeds to the next step.

(2) To obtain unbiased point estimate, only two patients are needed; one takes the placebo and the other drinks water at random. The unbiased estimate of treatment difference is given by $y - x$, where x and y are the responses from the placebo and water groups, respectively.

(3) To obtain the confidence interval and p-value, 4 patients are required: 2 for each group. The confidence interval is given by

$$CI = \delta \pm Z_{1-\alpha}\hat{\sigma},$$

where $\delta = \frac{y_1+y_2-x_1-x_2}{2}$ and $\hat{\sigma}^2 = var\,(\delta)$.
The p-value is given by

$$p = 1 - \Phi\left(\frac{\delta}{\hat{\sigma}}\right)$$

Here is the problem. The rejection of the null may not be consistent with the confidence interval and p-value. To be consistent, I can use 4 patients from Steps 1 to 3.

(4) Spring water is presumably very safe.

(5) It is a cost-effective approach. The water experiment may have a low power, only about 5%, but the cost is very low. Furthermore, when the water is "proved" to be efficacious statistically, the observed treatment difference is often very big because of small sample-size. A big observed treatment difference implies a big market value. Here is the overall picture: The experiment has an extremely low cost and low power, but it has potentially a big market value if the null hypothesis of no treatment effect is rejected.

What if all pharmaceutical companies take spring water as the test drug? If so, then 5% of them will claim efficacy of the water for some indications. What if the water is replaced by some relatively safe, but not efficacious compound? Does this discourage good science — just pick a safe compound and run a small trial? Is doing good science not a cost-effective approach?

To make an analogy, the spring water experiment can be repeated again and again until successful just as different companies can screen the same compounds again and again for the same or different indications until they find something. The water experiment is very easy to conduct and any

operational bias can virtually be avoided. In conclusion, don't overemphasize the importance of the type-I error rate, unbiased estimate, adjusted p-value, confidence interval, and the operational bias.

18.6 Summary

Adaptive designs violate several commonly accepted statistical principles. The violations on one hand remind us to adopt the new approaches with extreme caution. On the other hand, the new approaches may suggest that principles of inference only go so far and that new principles may be desirable. Indeed, adaptive designs present great challenges to frequentist statistics and conventional thinking in development. Group sequential designs have been widely used in clinical trials; however, there are many controversial issues with these designs that are only fully realized when we widen the concept to a more general category of adaptive designs (GSD can be viewed as an adaptive design with discrete values of sample-size adjustment, i.e., 0 increase if futility or efficacy, or add n_2 subjects otherwise). Those controversial issues include p-values and estimations. Should p-values and estimations be unconditional or conditional? Should they be conditional on statistical significance or stagewise-conditional? Should a test statistic be ordered by stage, by the mean, or by something else? The concepts of p-value and unbiasedness are based on repeated experiments, but what does this mean in clinical trials? The likelihood principle, which is equivalent to the conjunction of the frequentist sufficiency and the conditionality principles, denies the importance of hypothesis testing. All of these considerations seem to suggest that we should use other approaches, such as Bayesian approaches and decision theory. It is obvious that efficiency in clinical trials is not identical to the power of hypothesis testing. In clinical trials, the concept of efficiency often includes the power, the time to market, and the operationally flexibility. Over-emphasizing the power can be very misleading when evaluating adaptive designs. This simple fact is often overlooked.

One fundamental difficulty in using decision theory is that different beneficial bodies have different utilities; therefore, if a decision involves multiple decision-making bodies, the decision becomes extremely challenging. For example, if a regulatory body enforces the α criterion, we can use decision theory and incorporate the condition $p < \alpha$ as a constraint. This is not an ideal solution because $p < \alpha$ is not the best criterion, and the interactions between sponsors and the FDA will alter opinions.

We describe the deterministic world as a random phenomenon because

of our limited capabilities. Many controversial issues raised by the different statistical theories about this virtual random world can be reduced to a single, most important question: Should information level be dependent on the (experimental) procedure?

"From where we stand, the rain seems random. If we could stand somewhere else, we would see the order in it." — T. Hilberman

"The world is deterministic, just like a sequence of virtual random numbers generated by computer and by studying the latter, we can understand the statistical controversies." — Mark Chang

Appendix A

Random Number Generation

A.1 Random Number

To perform clinical trial simulations, we need to take random samples. Typically, random sampling is based on the computer-generated, uniformly distributed random numbers over (0,1). The computer-generated "random" number is not true random because the sequence of the numbers is determined by the so-called seed, an initial number. Other random variates from a nonuniform distribution are usually obtained by applying a transformation to uniform variates. There are usually several algorithms available to generate random numbers with a specific distribution. The algorithms differ in speed, in accuracy, and the memory required.

A.2 Uniformly Distributed Random Number

One of the commonly used methods to generate pseudorandom numbers starts with an initial value x_0, called the seed, and then recursively computes successive values $x_n, n \geq 1$, by letting

$$x_n = ax_{n-1} \quad module\ m, \tag{A2.1}$$

where a and m are given positive integers. (A2.1) means that ax_{n-1} is divided by m and the remainder is taken as the value of x_n. Thus, each x_n is either $0, 1, ...,$ or $m-1$ and the quantity x_n/m is called a pseudorandom number, which is approximately uniformly distributed on $(0, 1)$. This method is called linear congruential method. The positive integer a directly impacts the quality of the random deviates. m is the period of the sequence of the random numbers. the number a should be carefully chosen such that lead to a large m. Park and Miller (1988) has suggested a=$7^5 = 16807$, or $m = 2^{31} - 1 = 2147483647$. Please be aware that not all built-in random number generators from software products are good.

A.3 Inverse CDF Method

It is well-known that if X is a scalar random variable with a continuous cumulative distribution function (c.d.f.) F, then the random variable $U = F(X)$ has a U(0,1) distribution. Hence we have $X = F^{-1}(U)$. This fact provides the so-called inverse c.d.f. technique for generating random numbers with the distribution F by using the random numbers from the uniform distribution. The inverse c.d.f. relationship exists between any two continuous (nonsingular) random variables. If X is a continuous random variable with c.d.f. F and Y is a continuous random variable with c.d.f. G, then $X = F^{-1}(G(Y))$ over the ranges of positive support. Using this kind of relationship is actually to match percentile points, of one distribution (F) with those of another distribution (G). The advantages of the inverse method is simple. However, the closed form of F^{-1} is not always available. When F does not exist in closed form, the inverse c.d.f. method can be applied by solving the equation $F(x) - u = 0$ numerically.

The inverse c.d.f. method also applies to discrete distributions. Suppose the discrete random variable X has mass points of $m_1 < m_2 < m_3 < ...$with probabilities of $p_1, p_2, p_3, ...$,and the distribution function

$$F(x) = \sum_{i \in m_i \leq x} p_i. \tag{A2.2}$$

To use the inverse c.d.f. method for this distribution, we first generate a realization u of the uniform random variable U. We then deliver the realization of the target distribution as x, where x satisfies the relationship

$$F(x_{(-)}) < u \leq F(x). \tag{A2.3}$$

A.4 Acceptance-Rejection Methods

The acceptance-rejection method is another elegant method for random number generation. For generating realizations of a random variable X with a distribution f, the acceptance-rejection method makes use of realizations of another random variable Y with a simpler distribution of g. Further, the p.d.f. g can be scaled to majorize f, using some constant c, i.e, $c\,g(x) > f(x)$ for all x. To gain efficiency, the difference $\varepsilon = c\,g(x) - f(x) > 0$ should be small for all x. The density g is called the majorizing density and $c\,g$ is called the majorizing function.

Algorithm A.1 The Acceptance-Rejection Method to Convert Uniform Random Numbers

1. Generate y from the distribution with density function g.

2. Generate u from a uniform (0,1) distribution.
3. If $u < f(y)/cg(y)$, then take y as the desired realization;
   otherwise, return to step 1.

Unlike the inverse CDF method, the acceptance-rejection can apply immediately to multivariate random variables.

Most software packages have certain capabilities to generate different random variables. We can use them to generate specific random variables. SAS Macro A.1 is an example to use SAS random function for the exponential distribution to generate the mixed exponential distribution. The SAS variables are corresponding the formulation (13.4) with $n = 2$.

››SAS Macro A.1 Mixed Exponential Distribution››

```
%Macro RanVars(nObs=100, Lamda1=1, Lamda2=1.5, w1=0.6, w2=0.4);
Data RVars;
Drop iObs;
Do iObs=1 To &nObs;
  xMixEXP=&w1/&Lamda1*Ranexp(782)+&w2/&Lamda2*Ranexp(323);
  Output;
End;
Run;
Proc Print Data=RVars; Run;
%Mend RanVars;
```

‹‹SAS‹‹

The example of using this SAS macro is presented below.

››SAS››

```
TITLE "Random Variables";
%RanVars(nObs=10, Lamda1=1, Lamda2=1.5, w1=0.6, w2=0.4);
```

‹‹SAS‹‹

A.5 Multi-Variate Distribution

Many statistical software products such as SAS and ExpDesign Studio® provide built-in functions for generating variates of univariate random numbers. However, Some of them may not provide random number generator for multi-variate random numbers. Here we provide the method to generate random variables from multivariate normal distribution.

The multivariate normal distribution is given by

$$f(\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{n/2} |\mathbf{M}|^{-1/2} \exp\left(-\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{2}\right), \tag{A2.4}$$

where $\mathbf{M} = \{m_{kl}\}$ is the covariance matrix.

Let $\mathbf{u} = (u_1, ..., u_n)$ be n independent variables from the standard normal distribution and A be the triangle matrix

$$A = \begin{pmatrix} a_{11} & 0 & ... & 0 \\ a_{21} & a_{22} & .. & 0 \\ ... & ... & ... & ... \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}. \tag{A2.5}$$

Let $\boldsymbol{\xi} = \mathbf{A}\mathbf{u}$ or

$$\begin{cases} \xi_1 = a_{11}u_1 \\ \xi_2 = a_{21}u_1 + a_{22}u_2 \\ ..... \\ \xi_n = a_{n1}u_1 + a_{n2}u_2 + ... + a_{nn}u_n. \end{cases} \tag{A2.6}$$

It is obvious that the expectation $E(\xi_k) = 0$ $(k = 1, ..., n)$. Furthermore, we have

$$E(u_k u_l) = \delta_{kl} = \begin{cases} 1, & k = l \\ 0, & k \neq l. \end{cases} \tag{A2.7}$$

Therefore it can be obtained that

$$E(\xi_k \xi_l) = E\left(\left(\Sigma_{i=1}^{k} a_{ki}u_i\right)\left(\Sigma_{j=1}^{l} a_{kj}u_j\right)\right) = \Sigma_{j=1}^{k} a_{kj}a_{lj} = \left(\mathbf{A}\mathbf{A}^T\right)_{kl}. \tag{A2.8}$$

Comparing the corresponding elements in $\mathbf{A}\mathbf{A}^T = \mathbf{M}$, we can obtain:

$$a_{kl} = \frac{m_{kl} - \Sigma_{j=1}^{l-1} a_{kj}a_{lj}}{\sqrt{m_{ll} - \Sigma_{j=1}^{l-1} a_{lj}^2}}, \quad (1 \leq l \leq k \leq n). \tag{A2.9}$$

For convenience, we have defined $\Sigma_{j=1}^{0} a_{kj}a_{lj} = 0$.

Algorithm A.2 The Generation of Random Numbers ξ_i from Multivariate Normal Distribution

1. Calculate a_{kl} using (A2.9) for $1 \leq l \leq k \leq n$.

2. Generate n independent random numbers u_i $(i = 1, ..., n)$ from the standard normal distribution.

3. Generate ξ_i using (A2.6).

Algorithm A.2 is implemented in SAS Macro A.2, which can be used to generate the standard multivariate normal distribution. The SAS variables are defined as follows: **nVars** = number of variables, **sSize** = square of nVars, **nObs** = number of observations to be generated, ss{i} = covariate matrix, x{i} = the outputs multivariates.

››SAS Macro A.2 Multi-Variate Normal Distribution››

```
%Macro RanVarMNor(sSize=4, nVars=2, nObs=10);
Data nVars; SET CorrMatrix; keep x1-x&nVars;
Array a{&sSize}; Array xNor{&nVars}; Array x{&nVars};
 Array ss{&sSize}; * Correlation matrix;
* Checovsky decomposition;
 Do k=1 To &nVars; Do L=1 To k;
 Saa=0; Sa2=0;
       Do j=1 to L-1;
            Saa=Saa+a{&nVars*(k-1)+j}*a{&nVars*(L-1)+j};
            Sa2=Sa2+a{&nVars*(L-1)+j}*a{&nVars*(L-1)+j};
       End;
      nkL=&nVars*(k-1)+L;
 a{nkL}=(ss{nkL}-Saa)/Sqrt(ss{&nVars*(L-1)+L}-Sa2);
 End; End;
Do iObs=1 to &nObs;
 Do iVar=1 to &nVars; xNOR{iVar}=Rannor(762); End;
      Do iVar=1 to &nVars;
       x{iVar}=0;
       Do i=1 To iVar;
            x{iVar}=x{iVar}+a{&nVars*(iVar-1)+i}*xNor{i};
      End;
      End;
 Output;
End;
Run;
Proc Print data=nVars(obs=100); Run;
Proc corr data=nVars; Run;
%Mend RanVarMNor;
```

‹‹SAS‹‹

An example of SAS macro call to generate the standard multivariate

normal variables is given as follows:

```
>>SAS>>
TITLE "Standard Multivariate Normal Variables";
Data CorrMatrix;
 Array ss{9} (1, .1, .5,
              .1, 1, .5,
              .5, .5, 1);
%RanVarMNor(sSize=9, nVars=3, nObs=10000);
Run;
<<SAS<<
```

Using the standard multivariate normal variables, we can transform them into general multivariate normal variables as shown below.

```
>>SAS>>
Title "General Multivariate Normal Variables";
Data Final; Keep ys1-ys3;
 Set nVars;
 Array ys{3}; Array x{3};
 Array sigma{3} (1, 2, 3);
 Array means{3} (0, 1, 3);
 Do iVar=1 To 3;
      ys{iVar}=x{iVar}*sigma{iVar}+means{iVar};
 End;
 Title "Check the outputs";
Proc print data=Final; Run;
Proc means data=Final; Run;
Proc corr data=Final; Run;
<<SAS<<
```

Appendix B

Implementing Adaptive Designs in R

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. R compiler is available at http://www.r-project.org/.

Sample-Size Based on Conditional Power in Chapter 7

The sample-size required at the second stage based on the conditional power (7.15) is implemented in R Function B.1. The R variables are defined as follows: **nAdjModel** ="MIP", "MSP", "MPP", or "LW" for the four methods in Table 7.1: **alpha0** = overall α level; **alpha1** = efficacy stopping boundary at the first stage; **eSize** = standardized effect size; **cPower** = the conditional power; **p1** = the stagewise p-value at the first stage; **w1** and **w2** = weights for Lehmacher-Wassmer method; and **n2New** = new sample-size required for the second stage to achieve the desired conditional power.

>>R Function B.1: Sample-Size Based on Conditional Power>>>

```
nByCPower <- function(nAdjModel, a2, eSize, cPower, p1, w1, w2){
    if (nAdjModel=="MIP") {BFun <- qnorm(1-a2)}
    if (nAdjModel=="MSP") {
        BFun <- qnorm(1-max(0.000001,a2-p1))
    }
    if (nAdjModel=="MPP") {BFun <-qnorm(1- a2/p1)}
    if (nAdjModel=="LW") {
        BFun <- (qnorm(1-a2)- w1*qnorm(1-p1))/w2
    }
    2*((BFun-qnorm(1-cPower))/eSize)^2
```

```
}
```
‹‹R‹‹

››R Example››
```
pw1=nByCPower("MSP",.1,.3,.4,.05,.6,.4)
pw1
```
‹‹R‹‹

Sample-Size Re-Estimation in Chapter 9

R-function B.2 TwoArmNStgAdpDsg is developed to simulate a two-arm n-stage adaptive design with a normal, binary, or survival endpoint using "MIP", "MSP", "MPP", or "LW". The sample-size adjustment can be "CHW" based on (9.4) or the conditional power method (9.6). The sample-size adjustment is allowed only at the first interim analysis, and the sample-size adjustment affects only the final stagewise sample-size, **ux** and **uy** = the means, response rates, or hazard rates for the two groups, **Ns[k]** = sample-size per group at the stage k. **nMinIcr** = minimum sample-size increment for the conditional power approach only (this info is blinded to sponsor), **n2new** = the reestimated sample-size per group at the second stage, and **eSize** = the standardized effect size. **nSims** = number of simulation runs, **nStgs** = number of stages, alpha0 = overall α; and **EP** = "normal", "binary", or "survival". **Model** = "MSP", "MSP", "MPP," or "LW", **Nadj** = "N" for the case without SSR, **Nadj** = "Y" for the case with SSR; and **nAdjModel** = "MIP", "MSP", "MPP", or "CHW" for SSR based on the conditional power. **cPower** = conditional power, **DuHa** =1, **Nmax** = the maximum sample-size allowed, **N0** = the initial sample-size, **sigma** = standard deviation for normal endpoint, **tAcr** = accrual time, **tStd** = study duration, **power** = initial target power for the trial. **Aveux**, **Aveuy**, and **AveN** = average simulated responses (mean, proportion, or hazard rate) and sample size, **FSP[i]** = futility stopping probability at the i^{th} stage, **ESP[i]** = efficacy stopping probability at the i^{th} stage, **alpha[i]** and **beta[i]** = efficacy and futility stopping boundaries at the i^{th} stage.

››R Function B.2: Sample-Size Re-Estimation››
```
TwoArmNStgAdpDsg <- function (nSims=1000000, nStgs=2, ux=0, uy=1,
NId=0, a2=0.025, EP="normal", Model="MSP", Nadj="Y",
 nAdjModel="MSP", cPower=0.9, DuHa=1, Nmax=300, N0=200,
 nMinIcr=1, sigma=3, tAcr=10, tStd=24, Ns, alpha, beta) {
power=0; AveN=0; Aveux=0; Aveuy=0; cumN=0
for (i in 1:nStgs-1) {cumN=cumN+Ns[i]}
```

```
for (k in 1:nStgs) {
       sumWs[k]=0
       for (i in 1:k) {sumWs[k]=sumWs[k]+Ws[i]^2}
      sumWs[k]=sqrt(sumWs[k])
}
 u=(ux+uy)/2
 if (EP=="normal") {sigma=sigma}
 if (EP=="binary") {sigma=(u*(1-u))^0.5}
 if (EP=="survival") {
  expTerm=exp(-u*tStd)*(1-exp(u*tAcr))/(tAcr*u)
   sigma=u*(1+expTerm)^(-0.5)
}

for (i in 1:nStgs) { FSP[i]=0; ESP[i]=0 }
 for (iSim in 1:nSims) {
 ThisN=0;  Thisux=0;  Thisuy=0
 for (i in 1:nStgs) {TSc[i]=0}
 TS=0
 if (Model=="MPP") {TS=1}
 EarlyStop=0
 for (i in 1:nStgs) {
          uxObs=rnorm(1)*sigma/sqrt(Ns[i])+ux
          uyObs=rnorm(1)*sigma/sqrt(Ns[i])+uy
          Thisux=Thisux+uxObs*Ns[i]
          Thisuy=Thisuy+uyObs*Ns[i]
          ThisN=ThisN+Ns[i]
          TS0 = (uyObs-uxObs+NId)*sqrt(Ns[i]/2)/sigma
          if (Model=="MIP") {TS=1-pnorm(TS0)}
          if (Model=="MSP") {TS=TS+(1-pnorm(TS0))}
          if (Model=="MPP") {TS=TS*(1-pnorm(TS0))}
          if (Model=="LW") {
               for (k in 1:nStgs) {
                TSc[k]=TSc[k]+Ws[i]/sumWs[k]*TS0
               }
               TS=1-pnorm(TSc[i])
          }
          if (Model=="UWZ") {
              nT=(Thisuy-Thisux)/ThisN+NId
               TS0=nT*sqrt(ThisN/2)/sigma
               TS=1-pnorm(TS0)
          }
```

```
        if (TS>beta[i]) { FSP[i]=FSP[i]+1/nSims; EarlyStop=1 }
        else if (TS<=alpha[i]) {
              power=power+1/nSims
             ESP[i]=ESP[i]+1/nSims
              EarlyStop=1
        }
        else if (i==1 & Nadj=="Y") {
              eSize=DuHa/(abs(uyObs-uxObs)+0.0000001)
         nFinal=min(Nmax, max(N0,eSize*N0));

              if (nAdjModel != "CHW") {
              eSize=(uyObs-uxObs+NId)/sigma;
              n2New=nByCPower(nAdjModel, a2, eSize,
                          cPower, TS, ws[1], ws[2]);
             mT=min(Nmax,Ns[1]+n2New+nMinIcr/2)
              nFinal=round(mT, nMinIcr)
              }
              if (nStgs>1) {Ns[nStgs]= max(1,nFinal-cumN)}
        }
        if (EarlyStop==1) i=nStgs+1
 }

Aveux=Aveux+Thisux/ThisN/nSims
Aveuy=Aveuy+Thisuy/ThisN/nSims
AveN=AveN+ThisN/nSims
}
power=round(power,3);  AveN=round(AveN)
Aveux=round(Aveux,4); Aveuy=round(Aveuy,4)
FSP=round(FSP,3); ESP=round(ESP,3)
return (cbind(Model, power, Aveux, Aveuy, AveN,
          FSP, ESP, alpha, beta, cPower))
}
```

‹‹R‹‹

››R Example››

```
Ns <- c(100,100); alpha <- c(0.005, 0.205); beta <- c(0.25, 0.205)
Ws <- c(1,1); ESP <- c(0,0); FSP <- c(0,0); sumWs <- c(0,0);
TSc <- c(0,0)
AD1=TwoArmNStgAdpDsg(nSims=100000, nStgs=2,
 ux=0, uy=0, NId=0, a2=0.025, EP="normal",
 Model="MSP", Nadj="N", nAdjModel="MSP",
```

```
 cPower=0.9, DuHa=1, Nmax=300, N0=200,
 nMinIcr=1, sigma=3, tAcr=10, tStd=24, Ns, alpha, beta)

AD2=TwoArmNStgAdpDsg(nSims=100000, nStgs=2,
 ux=0, uy=1, NId=0, a2=0.025, EP="normal",
 Model="MSP", Nadj="Y", nAdjModel="MSP",
 cPower=0.9, DuHa=1, Nmax=300, N0=200,
 nMinIcr=1, sigma=3, tAcr=10, tStd=24, Ns, alpha, beta)

AD1; AD2
```

‹‹**R**‹‹

Drop-Loser Design in Chapter 11

R-function B.3, DrpLsrNRst, can be used to simulate the trial with drop-loser design using either weak or strong alpha-control. The weak control only controls α under the global null hypothesis. For the strong control, α is controlled under all null confutations. At the first stage, Bonferroni adjustment is used for the strong control by inflating the p-value from p_1 to (nArms-1)$p_{1\,\min}$, where $p_{1\,\min}$ is the smallest p-value among all the comparisons at the first stage. For the weak control, the first stage p-value p_1 is from a contrast test (see Chapter 2) and no p-value adjustment is required. The overall α is controlled by using MSP. The SAS variables are defined as follows. **nArms** = number of arms in the trial, **us[i]** = the true response (mean, rate, and hazard rate) in the i^{th} arm, **sigma** = common standard deviation, **N** = sample per group, **cPower** = the target conditional power at the interim analysis, **AveN** = average total sample-size, **Alpha1**= early efficacy stopping boundary (one-sided), **beta1**= early futility stopping boundary, **Alpha2** = the final efficacy stopping boundary. For the strong control, **CntlType** = "strong"; otherwise, the weak control is used; **NId** = noninferiority margin. The first arm must be the control arm.

››R Function B.3: Drop-Loser Design ››

```
DrpLsrNRst <- function(nSims=100000, CntlType="strong",  nArms=5,
 alpha=0.025, beta=0.2, NId=0, cPower=0.9, nInterim=50, Nmax=150,
 nAdj="Y", alpha1=0.01, beta1=0.15, alpha2=0.1871, EP="normal",
 sigma=1, tStd=24, tAcr=10, us, cs) {

u1 <- c(1); u2 <- c(1)
n1=nInterim; FSP=0; ESP=0; AveN=0; Power=0
if (EP=="binary") {sigma=(us[1]*(1-us[1]))^0.5}
```

```
if (EP=="survival") {
    expterm=exp(-us[1]*tStd)*(1-exp(us[1]*tAcr))
    sigma=us[1]*(1+expterm/(tAcr*us[1]))^(-0.5)
}
for (isim in 1:nSims) {
    TotalN=nArms*n1
    uMax=us[1]; Cntrst=0; SumSqc=0
    for (i in 1:nArms) {
        u1[i]=rnorm(1)*sigma/sqrt(n1)+us[i]
        if (u1[i]>uMax) { uMax=u1[i]; iMax=i }
        Cntrst=Cntrst+cs[i]*u1[i]
        SumSqc=SumSqc+cs[i]*cs[i]
      }
    Z1 = Cntrst*sqrt(n1)/sqrt(SumSqc)/sigma
    p1=1-pnorm(Z1)

    if (CntlType=="strong") {
        pNaive=(1-pnorm((uMax-us[1])/sigma*sqrt(n1/2)))
        p1=(nArms-1)*pNaive
    }
    if (p1>beta1) {FSP=FSP+1/nSims}
    if (p1<=alpha1) { Power=Power+1/nSims; ESP=ESP+1/nSims }

    if (iMax != 1 & p1>alpha1 & p1<=beta1) {
        BF=qnorm(1-max(0,alpha2-p1))-qnorm(1-cPower)
        n2=2*(sigma/(u1[iMax]-u1[1])*BF)^2
        nFinal=min(n1+n2, Nmax)
        if (nAdj=="N") {nFinal=Nmax}
        if (nFinal>n1) {
            TotalN=2*(nFinal-n1)+nArms*n1
            u2[1]=rnorm(1)*sigma/sqrt(nFinal-n1)+us[1]
            u2[iMax]=rnorm(1)*sigma/sqrt(nFinal-n1)+us[iMax]
            T2=(u2[iMax]-u2[1]+NId)*sqrt(nFinal-n1)/2^0.5/sigma
            p2=1-pnorm(T2); TS2=p1+p2
            if (TS2<=alpha2) {Power=Power+1/nSims}
        }
    }
AveN=AveN+TotalN/nSims
}
return (cbind(FSP, ESP, AveN, Power, cPower, Nmax))
}
```

‹‹**R**‹‹

```
>>R Example>>
 us <- c(.06, .12, .13, .14, .15)
 cs <- c(-0.54, .12, .13, .14, .15)

 D1=DrpLsrNRst(nSims=100000, CntlType="strong",
  nArms=5, alpha=0.025, beta=0.2, NId=0, cPower=0.9,
  nInterim=50, Nmax=150, nAdj="Y", alpha1=0.01,
  beta1=0.15, alpha2=0.1871, EP="normal", sigma=0.18,
  tStd=24, tAcr=10, us, cs)

 D2=DrpLsrNRst(nSims=100000, CntlType="strong",
  nArms=5, alpha=0.025, beta=0.2, NId=0, cPower=0.9,
   nInterim=50, Nmax=150, nAdj="Y", alpha1=0.01,
   beta1=0.15, alpha2=0.1871, EP="binary", sigma=0.18,
 tStd=24, tAcr=10, us, cs)

 us <- c(.06, .08, .09, .095, .095)

D3=DrpLsrNRst(nSims=100000, CntlType="strong",
  nArms=5, alpha=0.025, beta=0.2, NId=0, cPower=0.9,
  nInterim=50, Nmax=150, nAdj="Y", alpha1=0.01,
  beta1=0.15, alpha2=0.1871, EP="survival", sigma=0.18,
  tStd=24, tAcr=10, us, cs)

 D1;  D2;  D3
```

‹‹**R**‹‹

Biomarker-Adaptive Design in Chapter 12

R Function B.4 is developed for simulating biomarker-adaptive trials with two parallel groups. The key R variables are defined as follows: **Alpha1** = early efficacy stopping boundary (one-sided), **beta1** = early futility stopping boundary, **Alpha2** = final efficacy stopping boundary, **u0p** = response difference in biomarker-positive population, **u0n** = response in biomarker-negative population, **sigma** = asymptotic standard deviation for the response difference, assuming homogeneous variance among groups. For binary response, sigma $=\sqrt{r_1(1-r_1)+r_2(1-r_2)}$; For Normal response, sigma $=\sqrt{2}\sigma$. **np1**, **np2** = sample sizes per group for the first and second stage for the biomarker-positive population. **nn1**, **nn2** = sample sizes per

group for the first and second stage for the biomarker-negative population. **cntlType** = "strong", for the strong type-I error control and **cntlType** = "weak", for the weak type-I error control, **AveN** = average total sample-size (all arms combined), **pPower** = the probability of significance for biomarker-positive population, **oPower** = the probability of significance for overall population.

››R Function B.4: Biomarker-Adaptive Design‹‹

```
BMAD <- function(nSims=10, cntlType="strong", nStages=2, u0p=0.2,
 u0n=0.1, sigma=1, np1=50, np2=50, nn1=100,  nn2=100, alpha1=0.01,
 beta1=0.15,alpha2=0.1871) {
FSP=0; ESP=0; Power=0; AveN=0; pPower=0; oPower=0
 for (isim in 1:nSims) {
      up1=rnorm(1)*sigma/sqrt(np1)+u0p
      un1=rnorm(1)*sigma/sqrt(nn1)+u0n
      uo1=(up1*np1+un1*nn1)/(np1+nn1)
      Tp1=up1*sqrt(np1)/sigma
      To1=uo1*sqrt((np1+nn1))/sigma
      T1=max(Tp1,To1)
      p1=1-pnorm(T1)
      if (cntlType=="strong") {p1=2*p1}
      if (p1>beta1) {FSP=FSP+1/nSims}
      if (p1<=alpha1) {
           Power=Power+1/nSims; ESP=ESP+1/nSims
           if (Tp1>To1) {pPower=pPower+1/nSims}
           if (Tp1<=To1) {oPower=oPower+1/nSims}
      }
      AveN=AveN+2*(np1+nn1)/nSims
      if (nStages==2 & p1>alpha1 & p1<=beta1) {
           up2=rnorm(1)*sigma/sqrt(np2)+u0p
           un2=rnorm(1)*sigma/sqrt(nn2)+u0n
           uo2=(up2*np2+un2*nn2)/(np2+nn2)
           Tp2=up2*sqrt(np2)/sigma
           To2=uo2*sqrt(np2+nn2)/sigma
           if (Tp1>To1) {
                T2=Tp2
                AveN=AveN+2*np2/nSims
           }
           if (Tp1<=To1) {
                T2=To2
                AveN=AveN+2*(np2+nn2)/nSims
```

```
            }
            p2=1-pnorm(T2)
            TS=p1+p2
            if (TS<=alpha2) {
                Power=Power+1/nSims
                if (Tp1>To1) {pPower=pPower+1/nSims}
                if (Tp1<=To1) {oPower=oPower+1/nSims}
            }
        }
    }
return (cbind(FSP, ESP, Power, AveN, pPower, oPower))
}
```

‹‹**R**‹‹

››**R Example**››

```
# Simulation under global Ho, 2-stage design
bmad1=BMAD(nSims=100000, cntlType="strong", nStages=2, u0p=0,
      u0n=0, sigma=1.414, np1=260, np2=260, nn1=520, nn2=520,
      alpha1=0.01, beta1=0.15,alpha2=0.1871)

# Simulations under Ha, single-stage design
bmad2=BMAD(nSims=100000, cntlType="strong", nStages=1, u0p=0.2,
u0n=0.1, sigma=1.414, np1=400, np2=0, nn1=800, nn2=0, alpha1=0.025);

# Simulation under global Ho, 2-stage design
bmad3=BMAD(nSims=100000, cntlType="strong", nStages=2, u0p=0,
      u0n=0, sigma=1.414, np1=260, np2=260, nn1=520, nn2=520,
      alpha1=0.01, beta1=0.15,alpha2=0.1871)

# Simulations under Ha, 2-stage design
bmad4=BMAD(nSims=100000, cntlType="strong", nStages=2, u0p=0.2,
      u0n=0.1, sigma=1.414, np1=260, np2=260, nn1=520, nn2=520,
      alpha1=0.01, beta1=0.15,alpha2=0.1871)
bmad1; bmad2; bmad3; bmad4
```

‹‹**R**‹‹

Randomized Play-the-Winner Design in Chapter 14

R Function B.5 is developed to simulate RPW designs. The variables are defined as follows: the initial numbers of balls in the urn are denoted by **a0** and **b0**. Next **a1** or **b1** balls added to the urn if a response is observed in arm A or arm B. The SAS variables are defined as follows: **RR1**, **RR2**

= the response rates in group 1 and 2, respectively, **nSbjs** = total number of subjects (two groups combined), **nMin** (>0) = the minimum sample-size per group required to avoid an extreme imbalance situation, **nAnlys** = number of analyses (approximately an equal information-time design). All interim analyses are designed for randomization adjustment and only the final analysis for hypothesis testing. **aveP1** and **aveP2** = the average response rates in group 1 and 2, respectively. **Power** = probability of the test statistic > **Zc**. Note: **Zc** = function of (**nSbjs**, **nAnlys**, **a0**, **b0**, **a1**, **b1**, **nMin**).

>>R Function B.5: Randomized Play-the-Winner Design>>

```
RPW <- function(nSims=1000, Zc=1.96, nSbjs=200, nAnlys=3,
     RR1=0.2, RR2=0.3, a0=1, b0=1, a1=1, b1=1, nMin=1) {

set.seed(21823)
Power=0; aveP1=0; aveP2=0; aveN1=0; aveN2=0
for (isim in 1:nSims) {
 nResp1=0;  nResp2=0; N1=0;  N2=0
 nMax=nSbjs-nMin
 a=a0; b=b0; r0=a/(a+b)
 for (iSbj in 1:nSbjs) {
      nIA=round(nSbjs/nAnlys)
      if (iSbj/nIA==round(iSbj/nIA)) {r0=a/(a+b)}
      if ((rbinom(1,1,r0)==1 & N1<nMax) | N2>=nMax) {
          N1=N1+1
          if (rbinom(1,1,RR1)==1) {nResp1=nResp1+1; a=a+a1}
      }
      else
      {
          N2=N2+1
          if (rbinom(1,1,RR2)==1) { nResp2=nResp2+1; b=b+b1 }
      }
 }

 aveN1=aveN1+N1/nSims; aveN2=aveN2+N2/nSims

 p1=nResp1/N1;  p2=nResp2/N2
 aveP1=aveP1+p1/nSims; aveP2=aveP2+p2/nSims
 sigma1=sqrt(p1*(1-p1)); sigma2=sqrt(p2*(1-p2))
 sumscf=sigma1^2/(N1/(N1+N2))+sigma2^2/(N2/(N1+N2))
 TS = (p2-p1)*sqrt((N1+N2)/sumscf)
```

```
 if (TS>Zc) {Power=Power+1/nSims}
}
return (cbind(nSbjs, aveN1, aveN2, aveP1, aveP2, Zc, Power))
}
```
‹‹R‹‹

››R Example››

```
rpw1=RPW(nSims=1000, Zc=1.96, nSbjs=200, nAnlys=200, RR1=0.4,
          RR2=0.4, a0=1, b0=1,a1=0, b1=0, nMin=1)
rpw2=RPW(nSims=1000, Zc=1.96, nSbjs=200, nAnlys=200, RR1=0.3,
          RR2=0.5, a0=1, b0=1,a1=0, b1=0, nMin=1)
rpw3=RPW(nSims=10000, Zc=2.035, nSbjs=200, nAnlys=5, RR1=0.4,
          RR2=0.4, a0=2, b0=2,a1=1, b1=1, nMin=1)
rpw4=RPW(nSims=1000, Zc=2.035, nSbjs=200, nAnlys=5, RR1=0.3,
          RR2=0.5, a0=2, b0=2,a1=1, b1=1, nMin=1)
rpw1; rpw2; rpw3; rpw4
```
‹‹R‹‹

Continual Reassessment Method in Chapter 15

R Function B.6 is developed to simulate the 3+3 traditional escalation. The R variables are defined as follows: **nSims** = number of simulation runs, **nLevels** = number of dose levels, **DeEs** = "true" means that it allows for dose deescalation, otherwise, it does not. **AveMTD** = average MTD, **AveNPts** = average number of patients per trial, **AveNRsps** = average number of responses in a trial.

››R Function B.6: Continual Reassessment Method››

```
CRM <- function(nSims=100, nPts=30, nLevels=10, b=100, aMin=0.1,
  aMax=0.3, MTRate=0.3, nIntPts=100) {
nPtsAt <- c(1); nRsps <- c(1); RR <- c(1)
DLTs=0; AveMTD=0; VarMTD=0
dx=(aMax-aMin)/nIntPts
for (iSim in 1:nSims) {
      for (i in 1:nLevels) { nPtsAt[i]=0; nRsps[i]=0 }
 iLevel=1
 for (iPtient in 1:nPts) {
       iLevel=min(iLevel, nLevels)
       Rate=RRo[iLevel]
       nPtsAt[iLevel]=nPtsAt[iLevel]+1
      r= rbinom(1, 1, Rate)
      nRsps[iLevel]=nRsps[iLevel]+r
```

```
# Posterior distribution of a
     c=0
     for (k in 1:nIntPts) {
          ak=aMin+k*dx; Rate=1/(1+b*exp(-ak*doses[iLevel]))
          if (r>0) {L=Rate}
          if (r <= 0) {L=1-Rate}
          g[k]=L*g[k]; c=c+g[k]*dx
     }
     for (k in 1:nIntPts) {g[k]=g[k]/c}
# Predict response rate and current MTD
    MTD=iLevel; MinDR=1
    for (i in 1:nLevels) {
      RR[i]=0
      for (k in 1:nIntPts) {
        ak=aMin+k*dx
        RR[i]= RR[i]+1/(1+b*exp(-ak*doses[i]))*g[k]*dx
      }
       DR=abs(MTRate-RR[i])
       if (DR <MinDR) { MinDR = DR; iLevel = i; MTD = i  }
      }
 }
 for (i in 1:nLevels) {DLTs=DLTs+nRsps[i]/nSims}
 AveMTD=AveMTD+MTD/nSims
 VarMTD=VarMTD+MTD^2/nSims
}
SdMTD=sqrt(VarMTD-AveMTD^2)
return (cbind(nPts, nLevels, AveMTD, SdMTD, DLTs))
}
```

‹‹**R**‹‹

››R Example››

```
g <- c(1); doses <- c(1)
RRo <- c(0.01,0.02,0.03,0.05,0.12,0.17,0.22,0.4)
 for (k in 1:100) {g[k]=1}
 for (i in 1:8) {doses[i]=i}
crm1=CRM(nSims=500, nPts=8, nLevels=8, b=150, aMin=0, aMax=3,
     MTRate=0.17)
crm2=CRM(nSims=500, nPts=16, nLevels=8, b=150, aMin=0, aMax=3,
     MTRate=0.17)
crm1; crm2
```

‹‹**R**‹‹

Bibliography

Ades A.E, et al., (2006). Bayesian methods for evidence synthesis in cost-effectiveness analysis. Pharmacoeconomics, 24 (1): 1-19.

Alonso, A, et al. (2006). A unifying approach for surrogate marker validation based on Prentice's criteria. Statist. Med. 25:205–221. Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2315.

Arbuck, S.G. (1996). Workshop on phase I study design. Annals of Oncology, 7, 567-573.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. Biometrics, 11: 375-86.

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

Atkinson, A.C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. Biometrika, 69, 61-67.

Atkinson, A.C. and Donev, A.N. (1992). Optimum Experimental Designs. Oxford University Press, New York, New York.

Babb, J. S. and Rogatko, A. (2001). Patient specific dosing in a cancer phase I clinical trial. Statistics in Medicine, 20, 2079-2090.

Babb, J., Rogatko, A. and Zacks. S. (1998). Cancer phase I clinical trials. Efficient dose escalation with overdose control. Statistics in Medicine, 17, 1103-1120.

Babb, J.S. and Rogatko, A. (2004). Bayesian methods for cancer phase I clinical trials, Advances in Clinical Trial Biostatistics, Nancy L. Geller (ed.), Marcel Dekker, New York, New York.

Bandyopadhyay, U. and Biswas A. (1997). Some sequential tests in clinical trials based on randomized play-the-winner rule. Calcutta. Stat. Assoc. Bull., 47, 67-89.

Banerjee, A. and Tsiatis, A.A. (2006). Adaptive two-stage designs in phase II clinical trials. Statistics in Medicine, In press.

Banik N, Kohne K, and Bauer P. (1996). On the power of Fisher's combination test for two stage sampling in the presence of nuisance parameters. Biometrical Journal, 38:25-37.

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

Bauer, P. and Kohne K. (1994). Evaluation of experiments with adaptive interim analysis. Biometrics, 1029-1041.

Bauer, P. and Konig, F. The reassessment of trial perspectives from interim data—a critical view. Statistics in Medicine, 2005, in press.

Bauer, P. and Rohmel, J. (1995). An adaptive method for establishing a dose-response relationship, Statist. Med., 14: 1595-1607.

Bauer P. (1989). Multistage testing with adaptive designs (with Discussion). Biometrie und Informatik in Medizin und Biologie, 20:130-148.

Bauer, P. and Brannath, W. (2004). The advantages and disadvantages of adaptive designs for clinical trials. Drug Discovery Today 9, 351-57.

Bauer, P. and Kieser, M. (1999). Combining different phases in development of medical treatments within a single trial. Statistics in Medicine, 18, 1833-1848.

Bauer, P. and Köhne, K. (1996). Evaluation of experiments with adaptive interim analyses. Biometrics, 52, 380 (Correction).

Bauer, P. and König, F. (2006). The reassessment of trial perspectives from interim data – a critical view. Statistics in Medicine, 25, 23-36.

Bechhofer, R.E., Kiefer, J. and Sobel, M. (1968). Sequential Identification and Ranking Problems. University of Chicago Press, Chicago, Illinois.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289–300.

Berry, D.A. (2005). Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. Clinical Trials, 2: 295-300

Berry, D.A. (2005). Statistical innovations in cancer research. In: Holland J., et al., editors. Cancer Medicine. 7th ed. London: BC Decker, 411-25

Berry, D.A. and Eick, S.G. (1995). Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. Statistics in Medicine, 14, 231-246.

Berry, D.A. and Fristedt, B. (1985). Bandit problems: sequential allocation of experiments. Chapman Hall, London, UK.

Berry, D.A. and Stangl, D.K. (1996). Bayesian Biostatistics. Marcel Dekker, Inc. New York, New York.

Berry, D.A., et al., (2002). Adaptive Bayesian designs for dose-ranging drug trials. In Case Studies in Bayesian Statistics V. Lecture Notes in Statistics. Springer, New York, New York. 162 –181.

Birkett, N.J. (1985). Adaptive allocation in randomized controlled trials. Controlled Clinical Trials, 6, 146-155.

Birnbaum, A. (1962). On the foundations of statistical inference. Journal of America Statistical Association; Vol. 57, 269-306.

Bischoff, W. and Miller, F. (2005). Adaptive two-stage test procedures to find the best treatment in clinical trials. Biometrika, 92, 197-212.

Blackwell, D. and Hodges, J.L., Jr. (1957). Design for the control of selection bias. Annal of Mathematical Statistics, 28, 449-460.

Blume, J.D. (2002). Likelihood methods for measuring statistical evidence. Stat Med. 21: 2563-99

Brannath, W. et al., (2003). Sequential tests for non-inferiority and superiority. Biometrics, 59:106 –114. DOI: 10.1111=1541-0420.00013

Brannath, W., Konig, F. and Bauer, P. (2003). Improved repeated confidence bounds in trials with a maximal goal. Biometrical Journal, 45:311–324.

Brannath, W., Konig, F. and Bauer, P. (2006). Estimation in flexible two-stage designs. Statist Med. 2006; 25; 3366-3381.

Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests. J. of America Statistical Association; Vol. 97, No. 457, 236-244.

Brannath, W., and Bauer, P. (2004). Optimal conditional error functions for the control of conditional power. Biometrics 60, 715-723.

Branson, M. and Whitehead, W. (2002). Estimating a treatment effect in survival studies in which patients switch treatment. Statistics in Medicine, 21, 2449-2463.

Breslow, N.E. and Haug, C. (1977). Sequential comparison of exponential survival curves. JASA, 67, 691-697.

Bretz, F. et al. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. Biometrical Journal 48:4, 623–634 DOI: 10.1002 /bimj.200510232.

Bretz, F. and Hothorn, L.A. (2002). Detecting dose-response using contrasts: asymptotic power and sample-size determination for binary data. Statistics in Medicine, 21, 3325-3335.

Bronshtein, I.N. et al., (2004). Handbook of Mathematics. Springer-Verlag Berlin Heidelberg.

Brophy, J.M. and Joseph, L. (1995). Placing trials in context using Bayesian analysis. GUSTO revisited by reverend Bayes. Journal of American Medical Association, 273, 871-875.

Burman, C.F.and Sonesson, C. (2006). Are flexible designs sound? (with discussion) Biometrics 62, 664-683.

Buyse, M. and Molenberghs, G. Criteria for the validation of surrogate endpoints in randomized experiments. Biometrics 1998; 54:1014 –1029.

Buyse, M. et al. Statistical validation of surrogate endpoint. Drug Information Journal, 34, 49-67 & 447-454.

Campbell, M.J., Julious, S.A. and Altman, D.G. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. British Medical Journal 1995; 311: 1145-8.

Canner, P.L. Monitoring treatment differences in long-term clinical trials. Biometrics, 33, 603-615, 1997.

Chakravarty, A. (2005), Regulatory aspects in using surrogate markers in clinical trials. In: The evaluation of surrogate endpoint, Burzykowski, Molenberghs, and Buyse (eds.) 2005. Springer.

Chaloncr, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. Journal of Planning and Inference, 21, 191-208.

Chang, M. (2006). Adaptive design based on sum of stagewise p-values. Statist. Med. (in press). DOI: 10.1002/sim.2755.

Chang, M. (2007). Multiple-arm superiority and noninferiority designs with various endpoints. Pharmaceut. Statist. 6: 43-52. (www.interscience.wiley.com) DOI: 10.1002/pst.242

Chang, M., et al. (2007). BIO White Paper: Innovative approaches in drug develpment. Submitted.

Chang, M. (2005). Bayesian adaptive design with biomarkers. Presented at IBC's

Second Annual Conference on Implementing Adaptive Designs for Drug Development, November 7-8, 2005, Nassau Inn, Princeton, New Jersey.

Chang, M. (2006). Phase II/III seamless adaptive design. International Chinese Statistical Association Bulletin January, 42-47.

Chang, M. (2006). Bayesian Adaptive Design Method with Biomarkers. Biopharmaceutical Report. Volume 14, No. 2.

Chang, M. (2006). Recursive two-stage adaptive design, submitted.

Chang, M. (2007). Multiple-endpoint adaptive design, submitted.

Chang, M. (2005). Adaptive clinical trial design, in Pros. of the XIth International Symposium on Applied Stochastic Models and Data Analysis. Janssen, J. and Lenca, P. (Ed.), ISBN: 2-908849 -15 -1, Brest, France, ENST Bretagne, 2005.

Chang, M. (2006), Adaptive Design with Biomarkers, Conference on Innovating Clinical Drug Development, January 24-25, 2006, London, UK.

Chang, M. (2007), Clinical trial simulations in early development phases, in Encyclopedia of Biopharmaceutical Statistics, Chow, S.C. (Ed.). Taylor and Francis, New York, New York.

Chang, M. (2007), Clinical trial simulations in later development phases, in Encyclopedia of Biopharmaceutical Statistics, Chow, S.C. (Ed.). Taylor and Francis, New York, New York.

Chang, M. and Chow, S.C. (2005). A hybrid Bayesian adaptive design for dose response trials. Journal of Biopharmaceutical Statistics, 15, 667-691.

Chang, M. and Chow, S.C. (2006a). Power and sample-size for dose response studies. In Dose Finding in Drug Development. Ed. Ting, N. Springer, New York, New York.

Chang, M. and Chow, S.C. (2006b). An innovative approach in clinical development - utilization of adaptive design methods in clinical trials. Submitted.

Chang, M., Chow, S.C., and Pong, A. (2006). Adaptive design in clinical research - issues, opportunities, and recommendations. Journal of Biopharmaceutical Statistics, 16 No. 3, 299-309.

Chang, M.N. (1989). Confidence intervals for a normal mean following group sequential test. Biometrics, 45, 249-254.

Chang, M.N. and O'Brien, P.C. (1986). Confidence intervals following group sequential test. Controlled Clinical Trials, 7, 18-26.

Chang, M.N., Wieand, H.S., and Chang, V.T. (1989). The bias of the sample proportion following a group sequential phase II trial. Statistics in Medicine, 8, 563-570.

Chen, J.J., Tsong, Y. and Kang, S. (2000). Tests for equivalence or noninferiority between two proportions, Drug Information Journal, 34, 569-578.

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

Chen, T.T. and Ng, T.H. (1998). Optimal flexible designs in phase II cancer clinical trials. Statistics in Medicine, 17, 2301-2312.

Chen, Y.H.J., DeMets, D.L. and Lan, K.K.G. (2004). Increasing the sample-size when the unblinded interim result is promising. Statist. Med. 23:1023–1038 (DOI: 10.1002/sim.1688).

Cheng, Y. and Shen, Y. Estimation of a parameter and its exact confidence interval following sequential sample-size re-estimation trials. Biometrics 2004;

60:910 –918.

Chevret, S. (1993). The continual reassessment method in cancer phase I clinical trials: A simulation study. Statistics in Medicine, 12, 1093-1108.

Chevret, S. (Ed., 2006). Statistical methods for dose-finding experiments. John Wiley & Sons Ltd. West Sussex, England.

Chow, S.C., Chang, M. (2006). Adaptive Design Methods in Clinical Trials. Chapman & Hall/CRC, 2006.

Chow, S.C., Chang M, and Pong A. (2005). Statistical consideration of adaptive methods in clinical development. Journal of Biopharmaceutical Statistics, 15: 575-91

Chow, S.C. and Liu, J.P. (2003). Design and Analysis of Clinical Trials. Ed. 2. John Wiley & Sons: New York.

Chow, S.C. and Chang, M. (2005). Statistical consideration of adaptive methods in clinical development. Journal of Biopharmaceutical Statistics, 15; 575-591.

Chow, S.C. and Liu, J.P. (2003). Design and Analysis of Clinical Trials. 2nd edition, John Wiley & Sons, New York, New York.

Chow, S.C. and Shao, J. (2002). Statistics in Drug Research. Marcel Dekker, Inc., New York, New York.

Chow, S.C. and Shao, J. (2005). Inference for clinical trials with some protocol amendments. Journal of Biopharmaceutical Statistics, 15, 659-666.

Chow, S.C. and Shao, J. (2006). On margin and statistical test for noninferiority in active control trials. Statistics in Medicine, 25, 1101-1113.

Chow, S.C., Shao, J., and Hu, Y.P. (2002). Assessing sensitivity and similarity in bridging studies. Journal of Biopharmaceutical Statistics, 12, 385-400.

Chow, S.C., Shao, J., and Wang, H. (2003). Sample Size Calculation in Clinical Research. Marcel Dekker, Inc., New York, New York.

Chuang, S.C. and Agresti, A. (1997). A review of tests for detecting a monotone dose-response relationship with ordinal response data. Statistics in Medicine, 16: 2599-618.

Chuang, S.C. et al., (2006). Sample size re-estimation. Submitted.

Coad, D.S. and Rosenberger, W.F. (1999). A comparison of the randomized play-the-winner and the triangular test for clinical trials with binary responses. Statistics in Medicine, 18, 761-769.

Coburger, S. and Wassmer, G. (2001). Conditional point estimation in adaptive group sequential test designs. Biometrical Journal. 43:821–833.

Coburger, S. and Wassmer, G. (2003). Sample size reassessment in adaptive clinical trials using a bias corrected estimate. Biometrical Journal, 45:812–825.

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests. Biometrics, 10: 417-51.

Code of Federal Regulations, Title21, Section 312.21 Phases of an Investigation.

Cohen, A. and Sackrowitz, H.B. (1989). Exact tests that recover interblock information in balanced incomplete block design. Journal of American Statistical Association, 84, 556-559.

Collette, L. et al., Is Prostate-Specific Antigen a Valid Surrogate End Point for Survival in Hormonally Treated Patients with Metastatic Prostate Cancer? Joint Research of the European Organisation for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharma-

ceuticals. J Clin Oncol 2005; 23:6139-6148.
Conawav, M.R. and Petroni, G. R. (1996). Designs for phase II trials allowing for a trade-off between response and toxicity. Biometrics, 52, 1375-1386.
Conley, B.A. and Taube, S.E. (2004). Prognostic and predictive marker in cancer. Disease Markers 20, 35-43.
Cornfield, J. Recent methodological contributions to clinical trials. American Journal of Epidemiology 1976; 104 (4): 408-421
Cox, D.R. (1952). A note of the sequential estimation of means. Proc. Camb. Phil. Soc., 48, 447-450.
Cox, D.R. and Oakes, D. (1984). Analysis of Survival Data. Monographs on Statistics and Applied Probability. Chapman & Hall, London.
Cox, D.R. (2006). Principles of Statistical Inference. Cambridge University Press.Cambridge, UK.
Crowley, J. (2001). Handbook of Statistics in Clinical Oncology, Marcel Dekker, Inc., New York, New York.
Crowder, M. J. (2001). Classical Competing Risks, Chapman & Hall/CRC, Boca Raton.
CTriSoft Intl. (2002). Clinical Trial Design with ExpDesign Studio®, www.ctrisoft.net.
Cui, L., Hung, H.M.J., and Wang, S.J. (1999). Modification of sample-size in group sequential trials. Biometrics, 55, 853-857.
Dargie, H.J. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. Lancet 2001; 357: 1385 90.
De Gruttola, V.G., et al. Considerations in the Evaluation of Surrogate Endpoints in Clinical Trials: Summary of a National Institutes of Health Workshop. Controlled Clinical Trials 2001; 22:485–502.
DeMets, D. L. and Lan, K.K.G. (1994). Interim analysis – the alpha-spending function-approach. Statistics in Medicine 13, 1341-1352.
DeMets, D.L. and Ware, J.H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. Biometrika, 67, 651-660.
DeMets, D.L. and Ware, J.H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. Biometrika, 69, 661-663.
Denne, J.D. Sample size recalculation using conditional power. Statistics in Medicine 2001; 20:2645 –2660.
Denne, J. S. and Jennison, C. (2000). A group sequential t-test with updating of sample-size. Biometrika 87, 125-134.
Dent. S. F. and Fisenhauer, F. A. (1996). Phase I trial design: Are new methodologies being put into practice? Annals of Oncology, 6, 561-566.
Dmitrienko, A., et al. (2005), Analysis of Clinical Trials Using SAS. SAS Institute Inc., Cary. North Carolina.
Dmitrienko, A. and Wang, M.D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. Statist. Med. 2006; 25:2178-2195.
Dragalin, V. Adaptive designs: terminology and classification. Drug Inf J. (to appear).
Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50: 1096-121.

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

Efron, B. (1980). Discussion of "minimum chi-square, not maximum likelihood." Annal of Statistics, 8, 469-471.

Ellenberg, S.S., Fleming, T.R., and DeMets, D.L. (2002). Data Monitoring Committees in Clinical Trials – A Practical Perspective, John Wiley and Sons, New York, New York.

EMEA (2002). Point to Consider on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan. The European Agency for the Evaluation of Medicinal Products Evaluation of Medicines for Human Use. CPMP/EWP/2459/02, London, UK.

EMEA (2004). Point to Consider on the Choice of Non-inferiority Margin. The European Agency for the Evaluation of Medicinal Products Evaluation of Medicines for Human Use. London, UK.

EMEA (2006). Reflection Paper on Methodological Issues in Confirmatory Clinical Trials with Flexible Design and Analysis Plan. The European Agency for the Evaluation of Medicinal Products Evaluation of Medicines for Human Use. CPMP/EWP/2459/02, London, UK.

Emerson, S.S. and Fleming, T.R. Parameter estimation following group sequential hypothesis testing. Biometrika 1990; 77:875–892.

Ensign, L.G., et al. (1994). An optimal three-stage design for phase II clinical trials. Statistics in Medicine, 13, 1727-1736.

European Medicines Agency (EMEA). Committee for Medicinal Products for Human Use (CHMP). Guideline on the Evaluation of Anticancer Medicinal Products in Man. December 2005. Available from http:// www.emea.eu.int /pdfs/human/ewp/ 020595en.pdf. Date of access: 10 August 2006.

Fan, X. and DeMets, D.L. (2006). Conditional and unconditional confidence intervals following a group sequential test. Journal of Biopharmaceutical Statistics, 16: 107—122.

Fan, X., DeMets, D. L., and Lan, K.K.G. (2004). Conditional bias of point estimates following a group sequential test. J. Biopharm. Stat. 14:505-530.

Faries, D. (1994). Practical modifications of the continual reassessment method for phase I cancer clinical trials. Journal of Biopharmaceutical Statistics, 4, 147-164.

Farrington, C.P. and Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. Stat. Med. 9:1447-1454.

FDA (1988). Guideline for Format and Content of the Clinical and Statistical Sections of New Drug Applications. The United States Food and Drug Administration, Rockville, Maryland.

FDA (2000). Guidance for Clinical Trial Sponsors On the Establishment and Operation of Clinical Trial Data Monitoring Committees. The United States Food and Drug Administration, Rockville, Maryland.

FDA (2005). Guidance for Clinical Trial Sponsors. Establishment and Operation of Clinical Trial Data Monitoring Committees (Draft). Rockville, Maryland. http://www.fda.gov/ cber/qdlns/ clintrialdmc.htm.

FDA (March 2006). Innovation Stagnation, Critical Path Opportunities List. www.fda.gov

FDA Guidance for Industry (draft). Clinical Trial Endpoints for the Approval of Cancer Drug and Biologiecs. FDA, April, 2005. Available from URL: http://www.fda.gov/cder/Guidance/6592dft.htm. Date of access: 11 August 2006.

FDA. Draft Guidance for the use of Bayesian statistics in Medical Device Clinical Trials. www.fda.gov/cdrh/osb/guidance/1601.pdf. Accessed 22 May 2006.

FDA. Providing clinical evidence of effectiveness for human drug and biological products, guidance for industry [online]. Available from URL: http://www.fda.gov/cder/guidance/index.htm [Accessed 2005 July 20]

Fisher, L.D. and Moyé, L.A. Carvedilol and the Food and Drug Administration (FDA) approval process: an introduction. Control Clin Trials 1999; 20: 1-15

Fisher, L.D. (1999). Carvedilol and the Food and Drug Administration (FDA) approval process: the FDA paradigm and reflections on hypothesis testing. Control Clin Trials, 20: 16-39.

Fisher, L. D. (1998). Self-designing clinical trials. Statistics in Medicine 17, 1551–1562.

Fleming, T.R. and DeMets, D.L. (1996) Surrogate endpoint in clinical trials: are we being misled? Annals of internal medicine, 125, 605-613.

Follman, D.A., Proschan, M.A. and Geller, N.L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. Biometrics, 50, 325–336.

Freedman, L.S., Graubard. B.I. and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. Statistics in Medicine, 11:167–178.

Freidlin, B. and Simon, R. (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 2005;11(21) November 1, 2005.

Friede, T., et al. A comparison of procedures for adaptive choice of location tests in flexible two-stage designs. Biometrical Journal 2003; 45:292–310.

Friede, T. and Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A Review. Biometrical Journal 48, 537–555.

Friedman, B. (1949). A simple urn model. Comm. Pure Appl. Math., 2, 59-70.

Gallo, P. Confidentiality and trial integrity issues for adaptive designs, Drug Inf J. 2006 (submitted).

Gallo, P., et al. Adaptive Designs in Clinical Drug Development—An Executive Summary of the PhRMA Working Group. Journal of Biopharmaceutical Statistics 2006; 16: 275-83

Gasprini, M. and Eisele, J. (2000). A curve-free method for phase I clinical trials. Biometrics, 56, 609-615.

Gaydos, B., et al. Adaptive Dose Response Studies. Drug Inf J. 2006 (submitted).

Gelman, A., Carlin, J.B. and Rubin, D.B. (2003). Bayesian Data Analysis. 2nd Ed. Chapman & Hall/CRC. New York, New York.

Gilbert, N. Statistics. Philadelphia (PA): WB Saunders, 1976

Gillis, P.R. and Ratkowsky, D.A. (1978). The behaviour of estimators of the parameters of various yield-density relationships. Biometrics, 34, 191-198.

Gottlieb, S. (2006). Speech before 2006 Conference on Adaptive Trial Design, Washington, DC. http://www.fda.gov/ oc/speeches/ 2006/ trialdesign0710.html.

Goodman, S.N. (2005). Introduction to Bayesian methods I: measuring the strength of evidence. Clinical Trials, 2: 282-90

Goodman, S.N. (1999). Towards evidence-based medical statistics, I: the P-value fallacy. Annals of Internal Medicine, 130: 995—1004.

Goodman, S.N., Lahurak, M.L., and Piantadosi, S. (1995). Some practical improvements in the continual reassessment method for phase I studies. Statistics in Medicine, 5, 1149-1161.

Goodman, S.N. (2005). Introduction to Bayesian methods I: measuring the strength of evidence. Clinical Trials, 2, 282-290.

Gould, A.L. (1992). Interim analyses for monitoring clinical trials that do not maternally affect the type-I error rate. Statistics in Medicine, 11, 55-66.

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

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

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

Hallstron, A. and Davis, K. (1988). Imbalance in treatment assignments in stratified blocked randomization. Controlled Clinical Trials, 9, 375-382.

Hardwick, J.P. and Stout, Q. F. (1991). Bandit strategies for ethical sequential allocation. Computing Science and Stat., 23, 421-424.

Hardwick, J.P. and Stout, Q. F. (1993). Optimal allocation for estimating the product of two means. Computing Science and Stat., 24, 592-596.

Hardwick, J.P. and Stout, Q.F. (2002). Optimal few-stage designs. Journal of Statistical Planning and Inference, 104, 121-145.

Hartung, J. and Knapp G. A new class of completely self-designing clinical trials. Biometrical Journal 2003; 45:3–19.

Hartung, J. (2001). A self-designing rule for clinical trials with arbitrary response variables. Controlled Clinical Trials 22, 111–116.

Hartung, J. (2006). Flexible designs by adaptive plans of generalized Pocock- and O'Brien-Fleming-type and by self-designing clinical trials. Biometrical Journal 48, 521–536.

Hauben, M. and Reich, L. Safety related drug-labelling changes findings from two data mining algorithms. Drug Safety 2004; 27 (10): 735-44

Hauben, M., et al. (2005). Data mining in pharmacovigilance - the need for a balanced perspective. Drug Safety, 28 (10): 835-42

Hawkins, M. J. (1993). Early cancer clinical trials: safety, numbers, and consent. Journal of the National Cancer Institute, 85, 1618-1619.

Hedges, L.V. and Olkin, I. (1985). Statistical Methods for Meta-analysis. Academic Press, New York, New York.

Hellmich, M. (2001). Monitoring clinical trials with multiple arms. Biometrics, 57, 892-898.

Hochberg, Y. (1988). A sharper Bonferroni's procedure for multiple tests of significance. Biometrika, 75, 800-803.

Hochberg, Y. and Tamhane, A.C. (1987). Multiple comparison procedure. John Wiley & Son, Inc. New York, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scand.

J. Statist. 6, 65-70.

Holmgren, E.B. (1999). Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. Journal of Biopharmaceutical Statistics, 9, 651-659.

Hommel, G. 1988. A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75: 383–386.

Hommel G. Adaptive modifications of hypotheses after an interim analysis. Biometrical Journal 2001; 43: 581–589.

Hommel, G. and Kropf, S. (2001). Clinical trials with an adaptive choice of hypotheses. Drug Information Journal, 33, 1205-1218.

Hommel, G., Lindig, V. and Faldum, A. (2005). Two-stage adaptive designs with correlated test statistics. Journal of Biopharmaceutical Statistics, 15; 613-623.

Horwitz, R.I. and Horwitz, S.M. (1993). Adherence to treatment and health outcomes. Annals of Internal Medicine, 153, 1863–1868.

Hothorn, L. A. (2000). Evaluation of animal carcinogenicity studies: Cochran-Armitage trend test vs. Multiple contrast tests. Biometrical Journal, 42, 553-567.

Hsu, J. and Berger, R.L. Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. Journal of the American Statistical Association 1999; 94: 468-82.

Hughes, M.D. (1993). Stopping guidelines for clinical trials with multiple treatments. Statistics in Medicine, 12, 901-913.

Hughes, M.D. and Pocock, S.J. (1988). Stopping rules and estimation problems in clinical trials. Statistics in Medicine, 7, 1231-1242.

Hung, H. M. J., O'Neill, R. T., Wang, S. J. and Lawrence, J. (2006). A regulatory view on adaptive/flexible clinical trial design. Biometrical Journal, 48, 565–573.

Hung, H.M.J., et al. (2005). Adaptive statistical analysis following sample-size modification based on interim review of effect size. Journal of Biopharmaceutical Statistics, 15, 693-706.

Hung, H.M.J., et al. (2003). Some fundemental issues with non-inferiority testing in active controlled trials. Statistics in Medicine, 22, 213-225.

Hung, H. M. J. and Wang, S. J. (2004). Multiple testing of noninferiority hypotheses in active controlled trials. Journal of Biopharmaceutical Statistics 14, 327–335.

Hung, H. M. J., Wang, S. J., and O'Neill, R. (2006). Methodological issues with adaptation of clinical trial design. Pharmaceutical Statistics (to appear).

ICH (1996). International Conference on Harmonization Tripartite Guideline for Good Clinical Practice.

ICH E9 Expert Working Group (1999). Statistical principles for clinical trials (ICH Harmonized Tripartite Guideline E9). Statistics in Medicine, 18, 1905-1942.

Iglesias, C.P. and Claxton, K. (2006). Comprehensive Decision-analytic model and Bayesian value-of-information analysis. Pharmacoeconomics, 24 (5): 465-78

Inoue, L.Y.T., Thall, P.F. and Berry, D.A. (2002). Seamlessly expanding a randomized phase II trial to phase III. Biometrics, 58, 823–831.

Ivanova, A. and Flournoy, N. (2001). A birth and death urn for ternary outcomes: stochastic processes applied to urn models. In Probability and Statistical Models with Applications. Ed. Charalambides, C.A., Koutras, M.V., and Balakrishnan, N. Chapman and Hall/CRC Press, Boca Raton, Florida, 583-600.

Jeffery, H. (1961). Theory of probability. 3rd ed. Oxford: Oxford University Press.

Jennison, C. and Turnbull, B.W. Group Sequential Tests with Applications to Clinical Trials. Chapman & Hall: London/Boca Raton, Florida, 2000.

Jennison, C. and Turnbull, B.W. (1989). Interim analyses: the repeated confidence interval approach (with discussion). Journal of the Royal Statistical Society, Series B, 51:305 –361.

Jennison, C. and Turnbull, B.W. (2003). Mid-course sample-size modification in clinical trials. Statistics in Medicine, 22:971–993.

Jennison, C. and Turnbull, B.W. Repeated confidence intervals for group sequential clinical trials. Controlled Clinical Trials 1984; 5:33– 45.

Jennison, C. and Turnbull, B.W. (1990). Statistical approaches to interim monitoring of medical trials: a review and commentary. Statistics in Medicine, 5, 299-317.

Jennison, C. and Turnbull, B.W. (2005). Meta-analysis and adaptive group sequential design in the clinical development process. Journal of Biopharmaceutical Statistics, 15, 537-558.

Jennison, C. and Turnbull, B.W. (2006a). Adaptive and non-adaptive group sequential tests. Biometrika, 93, In press.

Jennison, C. and Turnbull, B.W. (2006b). Efficient group sequential designs when there are several effect sizes under consideration. Statistics in Medicine, 25, in press.

Jones, B. and Kenward, M. (2003). Design and Analysis of Cross-Over Trials, Second Edition. Chapman & Hall/CRC.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). Continuous univariate distributions, Vol. 1, John Wiley & Sons, New York, New York.

Julious, S.A. Tutorial in Biostatistics:Sample sizes for clinical trials with normal data. Statistics in Medicine 2004; 23: 1921-86.

Kalbeisch, J.D. and Prentice, R.T. (1980). The Statistical Analysis of Failure Time Data. Wiley, New York, New York.

Kalbfleisch, J.D. and R.L. Prentice (2002). The Statistical Analysis of Failure Time Data, 2nd edition, John Wiley & Sons, New York, New York.

Kelly, P., Stallard, N. and Todd, S. (2003). An adaptive group sequential design for clinical trials that involve treatment selection. Technical Report 03=1, School of Applied Statistics, The University of Reading.

Kelly, P.J., et al. (2005). A practical comparison of group-sequential and adaptive designs. Journal of Biopharmaceutical Statistics, 15, 719-738.

Kelly, P.J., Stallard, N. and Todd, S. (2005). An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. Journal of Biopharmaceutical Statistics, 15, 641-658.

Kieser, M. (2006). Inference on Multiple Endpoints in Clinical Trials with Adaptive Interim Analyses. Biometrical Journal. 41 3, 261-277

Kieser, M., Schneider, B. and Friede, T. A bootstrap procedure for adaptive selection of the test statistic in flexible two-stage designs Biometrical Journal

2002; 44:641– 652.

Kieser, M. and Friede, T. (2000). Re-calculating the sample-size in internal pilot study designs with control of the type-I error rate. Statistics in Medicine, 19, 901-911.

Kieser, M. and Friede, T. (2003). Simple procedures for blinded sample-size adjustment that do not affect the type-I error rate. Statistics in Medicine, 22, 3571-3581.

Kieser, M., Bauer, P., and Lehmacher, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. Biometrical Journal, 41, 261-277.

Kim, K. and DeMets, D.L. (1992). Sample size determination for group sequential clinical trial with immediate responses. Stats in Med. Vol 11., 1391-1399.

Kim, K. (1989). Point estimation following group sequential tests. Biometrics, 45, 613-617.

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

Kim, K. and DeMets, D. L. (1987). Confidence intervals following group sequential tests in clinical trials. Biometrics 43:857—864.

Kimko, H.C. and Duffull, S.B. (2003). Simulation for Designing Clinical Trials, Marcel Dekker, Inc., New York, New York.

Koch, A. (2006). Confirmatory clinical trials with an adaptive design. Biometrical Journal 48, 574–585.

Kokoska, S. and Zwillinger, D. (2000). Standard probability and statistics table and formulae, student edition. Chapman & Hall/CRC. Boca Raton, Florida.

Kramar, A., Lehecq, A., and Candalli, E. (1999). Continual reassessment methods in phase I trials of the combination of two drugs in oncology. Statistics in Medicine, 18, 1849-1864.

Lachin, J.M. and Foukes, M.A. (1986). Evaluation of sample-size and power for analysis of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics, 42: 507-19.

Lachin, J.M. (1988). Statistical properties of randomization in clinical trials. Controlled Clinical Trials, 9, 289-311.

Lan, K.K.G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. Biometrika, 70, 659–663.

Lan, K.K.G. and Wittes, J. (1988). The B-value: a tool for monitoring data. 44, 579-585.

Lan, K.K.G. and DeMets, D. L. (1989). Changing frequency of interim analysis in sequential monitoring. Biometrics 45, 1017–1020.

Lan, K.K.G. and Zucker, D. (1993). Sequential monitoring of clinical trials: The role of information in Brownian motion. Stat. Med. 12:753—765.

Lan, K.K.G (2002). Problems and issues in adaptive clinical trial design. Presented at New Jersey Chapter of the American Statistical Association, Piscataway, New Jersey, June 4, 2002.

Lan, K.K.G. and DeMets, D.L. (1987). Group sequential procedures: Calendar versus information time. Statistics in Medicine, 8, 1191-1198.

Shun, Z., Soo, Y. and Lan, K.K.G. (2007). Normal approximation in two-stage winner design. February 27, 2007. Dose Finding Clinical Trials Workshop. West Conshohocken, Pennsylvania.

Lang, T., Auterith, A. and Bauer, P. (2000). Trend tests with adaptive scoring. Biometrical Journal, 42:1007–1020.

Lawrence, J. and Hung, H.M.J. (2003). Estimation and confidence intervals after adjusting the maximum information. Biometrical Journal 45:143–152.

Lawrence, J. (2002). Strategies for changing the test statistics during a clinical trial. Journal of Biopharmaceutical Statistics, 12:193–205.

Lee, M.L.T., Chang, M. and Whitmore, G.A. (2007). A Threshold regression mixture model for assessing treatment efficacy in Multiple Myeloma Clinical Trial. (Working Manuscript).

Lee, M.-L.T., DeGruttola, V. and Schoenfeld, D. (2000). A model for markers and latent health status, J. Royal. Statist. Soc., Series B, 62, 747-762.

Lee M.-L. T. and Whitmore G. A. (2004). First hitting time models for lifetime data. In: Handbook of Statistics: Volume 23, Advances in Survival Analysis, C. R. Rao, N. Balakrishnan, editors, 537-543.

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

Lehmacher, W., Kieser, M. and Hothorn, L. (2000). Sequential and multiple testing for dose-response analysis. Drug Information Journal, 34, 591-597.

Lehmann, E.L. (1975). Nonparametric: Statistical Methods Based on Ranks. Holden-Day, San Francisco, California.

Lehmann, E.L. (1983). The Theory of Point Estimation. Wiley, New York, New York.

Lemuel, A. Moyé. (2003). Multiple analysis in clinical trials. Springer-Verlag, New York, Inc.

Li, H.I. and Lai, P.Y. (2003). Clinical trial simulation. In Encyclopedia of Biopharmaceutical Statistics, Ed. Chow, S.C., Marcel Dekker, Inc., New York, New York, 200-201.

Li, N. (2006). Adaptive trial design - FDA statistical reviewer's view. Presented at the CRT 2006 Workshop with the FDA, Arlington, Virginia, April 4, 2006.

Li, W.J., Shih, W.J. and Wang, Y. (2005), Two-stage adaptive design for clinical trials with survival data. Journal of Biopharmaceutical Statistics, 15; 707-718.

Lilford, R.J. and Braunholtz, D. (1996). For debate; The statistical basis of public policy: a paradigm shift is overdue. British Medical Journal, 313, 603-607.

Lin, Y. and Shih, W. J. (2001). Statistical properties of the traditional algorithm-based designs for phase I cancer clinical trials. Biostatistics, 2, 203-215.

Lindley, D.V. A statistical paradox. Biometrika, 1957; 44: 187-92

Liu, Q. and Chi, G.Y.H (2001). On sample-size and inference for two-stage adaptive designs. Biometrics; 57, 172-177.

Liu, Q. (1998). An order-directed score test for trend in ordered 2xK tables. Biometrics, 54, 1147-1154.

Liu, Q. and Pledger, G.W. (2005). Phase 2 and 3 combination designs to accelerate drug development. Journal of American Statistical Association, 100, 493-502.

Liu, Q., Proschan, M.A., and Pledger, G.W. (2002). A unified theory of two-stage adaptive designs. Journal of American Statistical Association, 97, 1034-1041.

Lokhnygina, Y. (2004). Topics in design and analysis of clinical trials. Ph.D. Thesis, Department of Statistics, North Carolina State Univeristy. Raleigh, North Carolina.

Louis, T.A. (2005). Introduction to Bayesian methods II: Fundamental concepts. Clinical Trials, 2: 291-94

Louis, T.A. (2005). Introduction to Bayesian methods II: Fundamental concepts. Clinical Trials, 2, 291-294.

Lynch, T.J, et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 350: 2129–39.

Maca, J., et al. (2006). Adaptive seamless phase II/III designs - background, operational aspects, and examples. Submitted.

Machin, D., et al. (1997). Statistical tables for the design of clinical studies. Ed. 2. Blackwell Scientific Publications: Oxford.

Maitournam, A. and Simon, R. On the efficiency of targeted clinical trials. Statistics in Medicine 2005; 24:329–339.

Marcus, R., Peritz, E. and Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63:655–660.

Marubini, E. and Valsecchi, M.G. (1995). Analysis Survival Data from Clinical Trials and Observational Studies, John Wiley & Sons, New York, New York.

Maxwell, C., Domenet, J.G., and Joyce, C.R.R. (1971). Instant experience in clinical trials: A novel aid to teaching by simulation. J. Clin. Pharmacol., 11, 323-331.

Mehta, C.R. and Patel, N.R. (2005). Adaptive, group sequential and decision theoretic approaches to sample-size determination. Submitted.

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

Melfi, V. and Page, C. (1998). Variability in adaptive designs for estimation of success probabilities. In New Developments and Applications in Experimental Design, IMS Lecture Notes Monograph Series, 34, 106-114.

Mendelhall, W. and Hader, R.J. (1985). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. Biometrika, 45, 504-520.

Molenberghs, G., Geys, H. and Buyse, M. Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. Statistics in Medicine 2001; 20:3023–3038

Molenberghs, G., Buyse, M. and Burzykowski, T. (2005). The history of surrogate endpoint validation, In The Evaluation of Surrogate Endpoint, Burzykowski, Molenberghs, and Buyse (eds.). Springer, N.Y.

Montori, V.M., et al. (2005). Randomized trials stopped early for benefit - a systematic review. Journal of American Medical Association, 294, 2203-2209.

Müller, H.H. and Shäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and classical group sequential approaches. Biometrics, 57, 886-891.

Müller, H.H. and Shäfer, H. (2004). A general ststistical principle of changing a

design any time during the course of a trial. Statist. Med., 23: 2497-2508.
Nam, J. (1987). A simple approximation for calculating sample sizes for detecting linear tend in proportions. Biometrics, 43: 701-5.
Neuhauser, M. (2001). An adaptive location-scale test. Biometrical Journal 43:809–819.
Neuhauser, M. and Hothorn, L. (1999). An exact Cochran-Armitage test for trend when dose-response shapes are a priori unknown. Computational Statistics & Data Analysis, 30, 403-412.
O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. Biometrika 35, 549–556.
Offen, W., et al. (2006). Multiple co-primary endpoints: Medical and statistical solutions. Drug Information Journal, In press.
Offen, W.W. (2003). Data Monitoring Committees (DMC). In: Encyclopedia of Biopharmaceutical Statistics. Ed. Chow, S.C., Marcel Dekker, Inc., New York, New York.
Ohman, J., Strickland, P. A. and Casella, G. (2003). Conditional inference following group sequential testing. Biometr. J. 45:515-526.
O'Neill, R.T. (1997). The Primary endpoint does not demonstrate clear statistical significance. Controlled Clinical Trials 18, 550-556.
O'Neill, R.T. (2004). A perspective on contributions of biostatistics to the critical path initiative. Presentation given at Workshop on "Model-based drug development – A cornerstone of the FDA's Critical Path Initiative", Basel Biometric Society, December 2004, available at URL: http://www.psycho.unibas.ch/BBS/slides/ONeill2.ppt.zip
O'Quigley, J. and Shen, L. (1996). Continual reassessment method: A likelihood approach. Biometrics, 52, 673-684.
O'Quigley, J., Pepe, M. and Fisher, L. (1990). Continual reassessment method: A practical design for phase I clinical trial in cancer. Biometrics, 46, 33-48.
Packer, M., et al. Effect of carvedilol on survival in severe chronic heart failure. N Eng J Med 2001; 344: 1651-58
Paez, J.G., et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science, 304: 1497–500.
Pahl, R., Ziegler, A., and Konig, I. K. (2006). Designing clinical trials using group sequential designs. Rnews 6, 21–26.
PAREXEL. (2003). PAREXEL's pharmaceutical R & D statistical sourcebook 2002/2003. Waltham, MA.
Park, S.K. and Miller, K.W. (1988). Communications of the ACM, vol 31, pp. 1192-1201.
Parmigiani, G. (2002). Modeling in medical decision making. John Wiley and Sons, West Sussex, England.
Paulson, E. (1964). A selection procedure for selecting the population with the largest mean from k normal populations. Annal of Mathematical Statistics, 35, 174–180.
Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. Biometrika. 64, 191–199.
Pocock, S.J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. Biometrics 38, 153–162.
Pocock, S.J. (2005). When (not) to stop a clinical trial for benefit. Journal of

American Medical Association, 294, 2228-2230.

Pocock, S.J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trials. Biometrics, 31, 103-115.

Pong, A. and Luo, Z. (2005). Adaptive design in clinical research. A special issue of the Journal of Biopharmaceutical Statistics, 15, No. 4.

Posch, M., et al. (2005). Testing and estimation in exible group sequential designs with adaptive treatment selection. Statist. Med. 24:3697–3714.

Posch, M., Bauer, P. and Brannath, W. Issues on flexible designs. Statistics in Medicine 2003; 22:953–969.

Posch, M., Bauer, P. Dealing with the unexpected: Modification of ongoing trials. Proceedings ROES Seminar St. Gallen, Switzerland, 2003.

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

Posch, M. and Bauer, P. (1999). Adaptive two-stage designs and the conditional error function. Biometrical Journal, 41, 689-696.

Posch, M., Bauer, P., and Brannath, W. (2003). Issues in designing flexible trials. Statistics in Medicine, 22, 953-969.

Posch, M., et al. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. Statistics in Medicine, 24, 3697-3714.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. Statistics in Medicine, 8:431– 440.

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

Proschan, M.A. and Wittes, J. (2000). An improved double sampling procedure based on the variance. Biometrics, 56, 1183-1187.

Proschan, M.A., Liu, Q.L. and Hunsberger, S. (2003). Practical midcourse sample-size modification in clinical trials. Controlled Clinical Trials 24:4-15.

Proschan, M.A. (2003). The geometry of two-stage tests. Statistica Sinica 13, 163–177.

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

Proschan, M.A., Follmann, D.A. and Waclawiw, M.A. (1992). Effects of assumption violations on type-I error rate in group sequential monitoring. Biometrics, 48, 1131-1143.

Proschan, M.A., Follmann, D.A. and Geller, N.L. (1994). Monitoring multiarmed trials. Statistics in Medicine, 13, 1441-1452.

Proschan, M.A., Leifer, E. and Liu, Q. (2005). Adaptive regression. Journal of Biopharmaceutical Statistics, 15, 593-603.

Proschan, M.A., Lan, K.K.G. and Wittes, J.T. (2006). Statistical Monitoring of Clinical Trials, a uniform approach. Springer, New York, New York.

Qu, Y. and Case, M. (2006). Quantifying the indirect treatment effect via surrogate markers. Statist. Med. 2006; 25:223–23. Published online 5 September 2005 in Wiley InterScience. DOI: 10.1002/sim.2176. URL: www.interscience.wiley.com.

Quinlan, J.A., Gallo, P., and Krams, M. (2006). Implementing adaptive designs:

logistical and operational consideration. Submitted.

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2005. www.R-project.org

Ravaris, C.L., et al. (1976). Multiple dose controlled study of phenelzine in depression anxiety states. Areh Gen Psychiatry, 33, 347-350.

Robert, C.P. (1997). The Bayesian Choice. Springer-Verlag New York, Inc.

Robins, J.M. and Tsiatis, A.A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics – Theory and Methods, 20, 2609–2631.

Rom, D.M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. Biometrika, 77, 663-665.

Rosenberger, W.F., et al. (2001). Optimal adaptive designs for binary response trials. Biometrics, 57, 909-913.

Rosenberger, W.F. and Lachin, J. (2002). Randomization in Clinical Trials, John Wiley and Sons, New York, New York.

Rosenberger, W.F. and Seshaiyer, P. (1997). Adaptive survival trials. Journal of Biopharmaceutical Statistics, 7, 617-624.

Rosner, G. L. and Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. Biometrika 75:723-729.

Royall, R. Statistical evidence: a likelihood paradigm. London: Chapman and Hall, 1997

Ruberg, S.J. (1998). Contrasts for identifying the minimum effective dose. J. of the American Statistical Association, 84: 816-22.

Ruberg, S.J. Dose response studies: I. Some design considerations. J. of Biopharmaceutical Statistics 1995; 5: 1-14.

Sampson, A. and Sill, M.W. (2005). Drop-the-losers Design: Normal Case. Biometrical Journal, 47, 257-268.

Sargent, D.J. and Goldberg, R.M. (2001). A flexible design for multiple armed screening trials. Statistics in Medicine, 20, 1051-1060.

Satagopan, J.M. and Elston, R.C. (2003). Optimal two-stage genotyping in population-based association studies. Genetic Epidemiology 25, 149–157.

Schafer, H. and Muller, H.H. (2001). Modification of the sample-size and the schedule of interim analyses in survival trials based on data inspections. Statistics in Medicine 20, 3741–3751.

Schafer, H., Timmesfeld, N. and Muller, H.H. (2006). An overview of statistical approaches for adaptive designs and design modifications. Biometrical Journal 48 (2006) 4, 507–520 DOI: 10.1002/bimj.200510234

Schaid, D.J., Wieand, S. and Therneau, T.M. Optimal two-stage screening designs for survival comparisons. Biometrika 1990; 77:507– 513.

Scherag, A., et al. (2003). Data adaptive interim modification of sample sizes for candidate-gene association studies. Human Heredity 56, 56–62.

Schmitz, N. (1993). Optimal Sequentially Planned Decision Procedures, Springer, New York, vol. 79 of Lecture Notes in Statistics.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of proportional hazards model. Biometrika 70:315—326.

Shaffer, J.P. (1995). Multiple hypothesis testing. Ann. Rev. Psychol. 1995. 46:561-84. Copyright © 1995 by Annual Reviews Inc.

Shao, J., Chang, M., and Chow, S.C. (2005). Statistical inference for cancer trials with treatment switching. Statistics in Medicine, 24, 1783-1790.

Shen, L. (2001). An improved method of evaluating drug effect in a multiple dose clinical trial. Statistics in Medicine, 20:1913 –1929. DOI: 10.1002=sim.842.

Shen, Y. and Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. Biometrics, 55, 190-197.

Shih, W.J. (2001). Sample size re-estimation - a journey for a decade. Statistics in Medicine, 20, 515-518.

Shih, W.J. and Lin, Y.(2006). Traditional and modified algorithm-based designs for phase I cancer clinical trials. in Cheveret S. (Ed., 2006). Statistical methods for dose-finding experiments. John Wiley & Sons. New York, New York.

Shirley, E. (1977). A non-parametric equivalent of William's test for contrasting increasing dose levels of treatment. Biometrics, 33, 386-389.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of American Statistical Association, 62, 626-633.

Siegmund, D. (1978). Estimation following sequential tests. Biometrika 65:341-349.

Siegmund, D. (1985). Sequential Analysis: Tests and Confidence Intervals. Springer. N.Y.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple test procedures. Journal of American Statistical Association, 81, 826-831.

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

Simon, R. (1979). Restricted randomization designs in clinical trials. Biometrics, 35, 503-512.

Simon, R. and Maitournam, A. (2004). Evaluating the efficiency of targeted designs for randomized clinical trials. Clinical Cancer Research. Vol. 10, 6759–6763.

Sommer, A. and Zeger, S.L. (1991). On estimating efficacy from clinical trials. Statistics in Medicine, 10, 45-52.

Spiegelhalter, D.J., Abrams, K.R. and Myles, J.P. Bayesian approach to clinical trials and health-care evaluation. Chichester: John Wiley & Sons, Ltd., 2004.

Stallard, N. and Rosenberger, W.F. (2002). Exact group-sequential designs for clinical trials with randomized play-the-winner allocation. Statist. Med. 21:467-480.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. Statistics in Medicine, 22, 689–703.

Stallard, N. and Todd, S. (2005). Point estimates and confidence regions for sequential trials involving selection. Journal of Statististical Planning and Inference, 135, 402-419.

Stewart, W. and Ruberg, S.J. (2000). Detecting dose response with contrasts. Statist Med 19: 913-21.

Strickland, P.A., Casella, G. (2003). Conditional inference following group sequential testing. Biometr. J. 45:515—526.

Susarla, V. and Pathala, K.S. (1965). A probability distribution for time of first

birth. Journal of Scientific Research, Banaras Hindu University, 16, 59-62.

Tang, D., Gnecco, C., and Geller, N.L. (1989). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. Biometrika 76:577– 583.

Tang, D. and Geller, N.L. (1999). Closed Testing Procedures for Group Sequential Clinical Trials with Multiple Endpoints. Biomatrics 55, 1188-1192.

Taves, D.R. (1974). Minimization - a new method of assessing patients and control groups. Clinical Pharmacol. Ther., 15, 443-453.

Taylor, H.M. and Karlin, S. (1998). An introduction to stochastic modeling. Academic Press Limited, London, UK.

Temple, R. (2005). How FDA currently make decisions on clinical studies. Clinical Trials 2: 276-81

Temple, R. (2006). FDA perspective on trials with interim efficacy evaluations. Statist. Med. (in press). DOI: 10.1002/sim.2631. URL: www.interscience.wiley.com.

Thach, C. T. and Fisher, L. D. (2002). Self-designing two-stage trials to minimize expected costs. Biometrics 58, 432-438.

Thall, P., Millikan, R. and Sung, H.G. (2000). Evaluating multiple treatment courses in clinical trials. Statisti. Med. 2000; 19:1011-1028.

Thall, P., Simon, R. and Ellenberg, S.S. (1989). A two-stage design for choosing among several experimental treatments and a control in clinical trials. Biometrics 45:537– 547.

Thall, P.F., Simon R, Ellenberg SS. (1988). Two-stage selection and testing designs for comparative clinical trials. Biometrika 75:303 –310.

The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). Harmonised Tripartite Guideline E4: Dose-Response Information to Support Drug Registration. 10 March 1994. URL: http://www.ich.org/LOB/media/ MEDIA480.pdf. Date of access: 10 August 2006.

Thomas, D.C., Haile, R.W. and Duggan, D. (2005). Recent developments in genomewide association scans: A workshop summary and review. American Journal of Human Genetics, 77, 337–345.

Thomas, D., Xie, R.R. and Gebregziabher, M. (2004). Two-stage sampling designs for gene association studies. Genetic Epidemiology 27, 401-414.

Tiemannn, R., Machnig, T. and Neuhaus, K.L. (2001). The Na+=H+ exchange inhibitor eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. Journal of the American College of Cardiology 38:1644-1651.

Timmesfeld, N., Schafer, H. and Muller, H. H. (2006). Increasing the sample-size during clinical trials with t-distributed test statistics without inflating the type-I error rate. In revision.

Ting, N. (2006) (Ed.). Dose Finding in drug development. Springer Science+Business Media Inc., N.Y.

Todd, S. (2003). An adaptive approach to implementing bivariate group sequential clinical trial designs. Journal of Biopharmaceutical Statistics, 13; No. 4, 605-619.

Todd, S. and Stallard, N. (2005). A new clinical trial design combining Phases 2 and 3: sequential designs with treatment selection and a change of endpoint.

Drug Information Journal, 39, 109-118.
Tonkens, R. (2005). An Overview of the Drug Development Process. May/June 2005 The Physicain Executive. p.51.
Tsiatis, A.A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. Biometrika; 90: 367-378.
Tsiatis, A.A., Rosner GL, and Metha, C.R. (1984). Exact confidence intervals following a group sequential test. Biometrics 40:797– 803.
Tukey, J.W. and Heyse, J.F. (1985). Testing the statistical certainty of a response to increasing doses of a drug, Biometrics 41: 295-301.
Uchida, T. (2006). Adaptive trial design - FDA view. Presented at the CRT 2006 Workshop with the FDA, Arlington, Virginia, April 4, 2006.
United States Department of Health and Human Services. Food and Drug Administration. Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. April 2005. Available from URL: http://www.fda.gov /cder/Guidance/6592dft.pdf. Date of access: 10 August 2006.
Wald, A. (1947). Sequential Analysis. Dover Publications, New York, New York.
Walton, M.K. (2006): PhRMA-FDA Adaptive Design Workshop.
Wang, S.J., et al. (2001). Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. Statistics in Medicine; 20:1903–1912.
Wang, S.J., Hung, H.M.J. and O'Neill, R.T. (2006). Adapting the sample-size planning of a phase III trial based on phase II data. Pharmaceut. Statist. 2006; 5: 85–97, Published online in Wiley InterScience, DOI: 10.1002/pst.217. URL: www.interscience.wiley.com.
Wang, Y. and McDermott, M.P. (1998). Conditional likelihood ratio test for a nonnegative normal mean vector. Journal of the American Statistical Association 93:380–386.
Wang, S.J. Regulatory experience of adaptive designs in well-controlled clinical trials. Presented at Adaptive Designs: Opportunities, Challenges and Scope in Drug Development, Washington, DC, Nov. 2006.
Wang, S.J. and Hung, H.M.J. (2005). Adaptive covariate adjustment in clinical trials. Journal of Biopharmaceutical Statistics, 15, 605-611.
Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for a sequential trials. Biometrics, 43, 193-200.
Wassmer, G. (1998), A Comparison of two methods for adaptive interim analyses in clinical trials. Biometrics 54, 696-705
Wassmer, G. (1999). Multistage adaptive test procedures based on Fisher's product criterion. Biometrical Journal 41, 279–293.
Wassmer, G. and Vandemeulebroecke, M. (2006). A brief review on software developments for group sequential and adaptive designs. Biometrical Journal 48, 732–737.
Wassmer, G., Eisebitt, R. and Coburger, S. (2001). Flexible interim analyses in clinical trials using multistage adaptive test designs. Drug information Journal, 35, 1131-1146.
Wei, L.J. (1977). A class of designs for sequential clinical trials. Journal of American Statistical Association, 72, 382-386.
Wei, L.J. (1978). The adaptive biased-coin design for sequential experiments.

Annal of Statistics, 9, 92-100.

Wei, L.J. and Durham, S. (1978). The randomized play-the-winner rule in medical trials. Journal of American Statistical Association, 73, 840-843.

Wei, L.J., et al. (1990). Statistical inference with data-dependent treatment allocation rules. JASA 85; 156-162.

Wei, L.J., Smythe, R.T. and Smith, R.L. (1986). K-treatment comparisons with restricted randomization rules in clinical trials. Annal of Statistics, 14, 265-274.

Weinthrau, M., et al. (1977). Piroxicam (CP 16171) in rheumatoid arthritis: A controlled clinical trial with novel assessment features. J Rheum, 4, 393-404.

Weir, C.J. and Walley, R.J. (2006). Statistical evaluation of biomarkers as surrogate endpoints: a literature review. Statist. Med. 2006; 25:183–203. Published online 26 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2319.

Westfall, P.H., et al. (1999), Multiple Comparisons and Multiple Tests, SAS Institute Inc., Cary, North Carolina.

White, I.R., et al. (1999). Randomisation-based methods for correcting for treatment changes: examples from the Concorde trial. Statistics in Medicine, 18, 2617–2634.

White, I.R., Walker, S, and Babiker, A.G. (2002). Randomisation-based efficacy estimator. Stata Journal, 2, 140–150.

Whitehead, J. (1993). Sample size calculation for ordered categorical data. Statistics in Medicine 12: 2257-71.

Whitehead, J. (1994). Sequential methods based on the boundaries approach for the clinical comparison of survival times (with discussions). Statistics in Medicine, 13, 1357-1368.

Whitehead, J. (1997). Bayesian decision procedures with application to dose-finding studies. International Journal of Pharmaceutical Medicine, 11, 201-208.

Whitehead, J. (1997). The Design and Analysis of Sequential Clinical Trials. Revised Second Edition. Chichester, UK: John Wiley.

Whitmore, G. A. (1983). A regression method for censored inverse-Gaussian data, Canadian. J. of Statist., 11, 305-315.

Whitmore, G. A., M. J. Crowder and J. F. Lawless (1998). Failure inference from a marker process based on a bivariate Wiener model, Lifetime Data Analysis, 4, 229-251.

Whitehead, J. and Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. Biometrics 39, 227–236.

Williams, D.A. A test for difference between treatment means when several dose levels are compared with a zero dose control. Biometrics1971; 27: 103-17.

Williams, D.A. (1972). Comparison of several dose levels with a zero dose control. Biometrics 28: 519-31.

Williams, G., Pazdur, R., and Temple, R. (2004). Assessing tumor-related signs and symptoms to support cancer drug approval. Journal of Biopharmaceutical Statistics, 14, 5-21.

Woodcock, J. FDA introductory comments: clinical studies design and evaluation issues. Clinical Trials 2005; 2: 273-75

Woodcock, J. (2004). FDA's Critical Path Initiative. URL: www.fda.gov/oc/

initia-tives/criticalpath/woodcock0602/woodcock0602.html

Zehetmayer, S., Bauer, P., and Posch, M. (2005). Two-stage designs for experiments with a large number of hypotheses. Bioinformatics, 21, 3771–3777.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. JASA, 1969; 80:974-984.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. Journal of Chronic Diseases, 28, 365-375.

Zucker, D.M., et al. (1999). Internal pilot studies II: comparison of various procedures. Statistics in Medicine, 19, 901-911.

Index

B

C

D

E

F

G

H

I

N

O

P

R

S

T

U

V

W

Z