ICSA Book Series in Statistics
Series Editors: Jiahua Chen · Ding-Geng (Din) Chen

Karl E. Peace
Ding-Geng Chen
Sandeep Menon *Editors*

Biopharmaceutical Applied Statistics Symposium

Volume 1 Design of Clinical Trials

Springer

ICSA Book Series in Statistics

More information about this series at http://www.springer.com/series/13402

Karl E. Peace · Ding-Geng Chen
Sandeep Menon
Editors

Biopharmaceutical Applied Statistics Symposium

Volume 1 Design of Clinical Trials

Springer

Editors
Karl E. Peace
Jiann-Ping Hsu College of Public Health
Georgia Southern University
Statesboro, GA, USA

Sandeep Menon
Boston University
Cambridge, MA, USA

Ding-Geng Chen
School of Social Work & Gillings School of Global Public Health
University of North Carolina
Chapel Hill, NC, USA

and

University of Pretoria
Pretoria, South Africa

ISSN 2199-0980 ISSN 2199-0999 (electronic)
ICSA Book Series in Statistics
ISBN 978-981-10-7828-6 ISBN 978-981-10-7829-3 (eBook)
https://doi.org/10.1007/978-981-10-7829-3

Library of Congress Control Number: 2017964432

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Currently, there are three volumes of the BASS Book Series, spanning 45 chapters. Chapters in this book are contributed by invited speakers at the annual meetings of the Biopharmaceutical Applied Statistics Symposium (BASS). Volume 1 is titled Design of Clinical Trials and consists of 15 chapters; Volume 2 is titled Biostatistical Analysis of Clinical Trials and consists of 12 chapters; and Volume 3 is titled Pharmaceutical Applications and consists of 18 chapters. The three volumes include the works of 70 authors or co-authors.

History of BASS: BASS was founded in 1994, by Dr. Karl E. Peace. Dr. Peace is the Georgia Cancer Coalition Distinguished Scholar/Scientist, Professor of Biostatistics, Founding Director of the Center for Biostatistics, and Senior Research Scientist in the Jiann-Ping College of Public Health at Georgia Southern University.

Originally, there were three objectives of BASS. Since the first editor founded the Journal of Biopharmaceutical Statistics (JBS) 3 years before founding BASS, one of the original objectives was to invite BASS Speakers to create papers from their BASS presentations and submit to JBS for review and publication. Ergo, BASS was to be a source of papers submitted to JBS to assist in the growth of the new journal JBS. The additional two objectives were:

- to provide a forum for pharmaceutical and medical researchers and regulators to share timely and pertinent information concerning the application of biostatistics in pharmaceutical environments; and most importantly,
- to provide revenues to support graduate fellowships in biostatistics at the Medical College of Virginia (MCV) and at the Jiann-Ping Hsu College of Public Health at Georgia Southern University (GSU).

After the JBS was on firm footing, the first objective was formally dropped. In addition, the third objective was expanded to include potentially any graduate program in biostatistics in the USA.

BASS I (1994) was held at the Hyatt Regency in Orlando, FL; BASS II–III were held at the Hilton Beach Resort, Inner Harbor, in San Diego, CA; BASS IV–VII were held at the Hilton Oceanfront Resort Hotel, Palmetto Dunes, in Hilton Head

Island, SC; BASS VIII–XII were held at the Desoto Hilton; and BASS XIII–XVI were held at the Mulberry Inn, both located in the Historic District of Savannah, GA. BASS XVII was held at the Hilton Resort Hotel at Palmetto Dunes, Hilton Head Island, SC. BASS XVIII–XIX were held at the Mulberry Inn in Savannah. To mark the twentieth Anniversary BASS meeting, BASS XX was held in Orlando at the Hilton Downtown Orlando Hotel. BASS XXI was held at the Holiday Inn Crowne Plaza in Rockville, MD; whereas BASS XXII and XXIII were held at the Radisson Hotel in Rockville, Maryland.

BASS XXIV (www.bassconference.org) was held at the Hotel Indigo in the charming historic Georgia city of Savannah. More than 360 tutorials and 57 1-day or 2-day short courses have been presented at BASS, by the world's leading authorities on applications of biostatistical methods attendant to the research, clinical development, and regulation of biopharmaceutical products. Presenters represent the biopharmaceutical industry, academia, and government, particularly the NIH and FDA.

BASS is regarded as one of the premier conferences in the world. It has served the statistical, biopharmaceutical, and medical research communities for the past 24 years by providing a forum for distinguished researchers and scholars in academia, government agencies, and industries to conduct knowledge sharing, idea exchange, and creative discussions of the most up-to-date innovative research and applications to medical and health care to enhance the health of general public, in addition to providing support for graduate students in their biostatistics studies. Toward this latter end, BASS has provided financial support for 75 students in completing their master's or doctorate degree in Biostatistics. In addition, BASS has provided numerous travel grants to doctorate-seeking students in Biostatistics to attend the annual BASS meeting. This provides a unique opportunity for students to broaden their education, particularly in the application of biostatistical design and analysis methods, as well as networking opportunities with biostatisticians from Academia, the Pharmaceutical Industry, and governmental agencies such as the FDA.

Volume 1 of the BASS Book Series, entitled Design of Clinical Trials, consists of 15 chapters. Chapter 1 presents statistical approaches to clinical trial simulations. Chapter 2 presents methods helpful in choosing the best function of baseline run-in data for use as a covariate in the analysis of treatment data from Phase III clinical trials in hypertension. Chapter 3 provides methods in designing adaptive trials in clinical research. Chapter 4 then provides best practices and recommendations for clinical trial simulations for adaptive designs. Chapter 5 discusses the design and analysis of clinical trials that collect recurrent event data.

Chapter 6 presents methods for response-adaptive allocation for binary outcomes in clinical trials from a Bayesian perspective. Chapter 7 addresses the important topic of high placebo response in neuroscience clinical trials. Chapter 8 presents methods for designing Phase I cancer clinical trials, for both single and combination agents. Chapter 9 discusses the structure for clinical trials that include sequential data monitoring procedures. Chapter 10 addresses both theory and practice in the design and data analysis of multiregional clinical trials. Chapter 11 continues

discussion of multiregional clinical trials with particular emphasis on ICH-E17 and subpopulations.

Chapter 12 also discusses multiregional clinical trials in the development of vaccines that are designed as adaptive group-sequential outcome studies. Chapter 13 deals with the development and validation of procedures for collecting patient-reported outcomes. Chapter 14 presents group-sequential and interim analysis and conditional power methods for survival trials from the nonproportional hazards perspective. Finally, Chap. 15 discusses the design and analysis of dose–response trials for early clinical development.

We are indebted to all the presenters, program committee, attendees, and volunteers who have contributed to the phenomenal success of BASS over its first 24 years, and to the publisher for expressing interest in and publishing the Series.

Statesboro, USA

Karl E. Peace, Ph.D.
Jiann-Ping Hsu College
of Public Health
Georgia Southern Univesity

Chapel Hill, USA/Pretoria, South Africa

Ding-Geng Chen, Ph.D.
Professor, University
of North Carolina
Extraordinary Professor
University of Pretoria

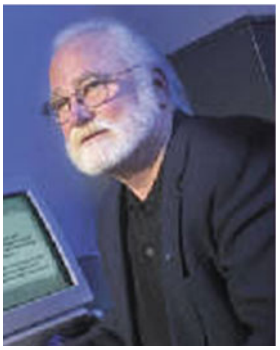
Cambridge, USA

Sandeep Menon
Vice President and Head
of Early Clinical Development
Biostatistics

Contents

Editors and Contributors

About the Editors



Prof. Karl E. Peace is currently Professor of Biostatistics, Senior Research Scientist, and Georgia Cancer Coalition Distinguished Scholar, in the Jiann-Ping Hsu College of Public Health (JPHCOPH) at Georgia Southern University, Statesboro, GA. He is a Fellow of the American Statistical Association (ASA), the Founding Editor of the Journal of Biopharmaceutical Statistics, Founding Director of the Center for Biostatistics in the JPHCOPH, Founder of BASS, and the Endower of JPHCOPH. He is the recipient of numerous awards and citations from the ASA, the Drug Information Association, the Philippine Statistical Association, BASS, and government bodies. He was cited by US and State of Georgia Houses of Representatives and the House of Delegates of Virginia for his contributions to Education, Public Health, Biostatistics, and Drug Research and Development. He is the author or editor of 15 books and over 100 publications.

Prof. Ding-Geng Chen is a Fellow of the American Statistical Association and currently the Wallace H. Kuralt Distinguished Professor at the University of North Carolina at Chapel Hill, USA, and an Extra-ordinary Professor at University of Pretoria, South Africa. He was a Professor at the University of Rochester and the Karl E. Peace endowed eminent scholar chair in Biostatistics at the Department of Biostatistics, Jiann-Ping Hsu College of Public Health, Georgia Southern University. He is also a senior consultant for biopharmaceuticals and government agencies with extensive expertise in clinical trial biostatistics and public health statistics. He has written more than 150 refereed publications and co-authored/co-edited 12 books on clinical trial methodology, meta-analysis, causal inference, and public health statistics.

Dr. Sandeep Menon is currently the Vice President and the Head of Early Clinical Development Statistics at Pfizer Inc. and also holds adjunct faculty positions at Boston University, Tufts University School of Medicine and Indian Institute of Management (IIM). He is the elected fellow of American Statistical Association. He is internationally known for his technical expertise especially in the area of adaptive designs, personalized medicine, multiregional trials, and small populations. He has co-authored and co-edited books and contributed to influential papers in this area.

He is the Vice Chair of Cross Industry/FDA-Adaptive Design Scientific Working Group under DIA (Drug Information Association); in the program committee for BASS and ISBS; and is in the advisory board for the M.S. in Biostatistics program at Boston University. He is serving as an associate editor of American Statistical Association (ASA) journal *Statistics in Biopharmaceutical Research* (SBR) and as a selection committee member of *Samuel S. Wilks Memorial Award* offered by ASA.

Contributors

Bruce Binkowitz Biometrics, Shionogi, Inc., Florham Park, NJ, USA

Joseph C. Cappelleri Pfizer Inc, Groton, CT, USA

Chi-Tian Chen Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, Republic of China

Fei Chen Janssen R & D, Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA

Shein-Chung Chow Duke University School of Medicine, Durham, NC, USA

Dave L. DeMets Department of Biostatistics, Wake Forest School of Medicine, Winston-Salem, NC, USA; Department Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

Gheorghe Doros Department of Biostatistics, Boston University, Boston, MA, USA

Sheri E. Fehnel RTI Health Solutions, Research Triangle Park, NC, USA

Yi Hao Frontier Science and Technology Research, Madison, WI, USA

Chin-Fu Hsiao Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, Republic of China

Edward Lakatos BiostatHaven, Inc., Croton-on-Hudson, USA

K. K. G. Lan Janssen R & D, Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA

Gang Li Janssen R & D, Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA

Pilar Lim Department of Quantitative Sciences, Janssen Research & Development, LLC, Titusville, NJ, USA

Jung-Tzu Liu Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan, Republic of China

Yuyin Liu Department of Quantitative Sciences, Janssen Research & Development, LLC, Titusville, NJ, USA

Cristiana Mayer Quantitative Sciences, Janssen Research & Development LLC, Titusville, NJ, USA

Lori D. McLeod RTI Health Solutions, Research Triangle Park, NC, USA

Stephan Ogenstad Statogen Consulting LLC, Wake Forest, NC, USA

Karl E. Peace Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA

Inna Perevozskaya Statistical Innovation Group, GSK, Collegeville, PA, India

Annpey Pong Merck Research Laboratories, Rahway, NJ, USA

David M. Reboussin Department of Biostatistics, Wake Forest School of Medicine, Winston-Salem, NC, USA

Roy T. Sabo Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

Yoko Tanaka Santen Inc., Emeryville, CA, USA

Hsiao-Hui Tsou Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, Republic of China; Graduate Institute of Biostatistics, College of Public Health, China Medical University, Taichung, Taiwan, Republic of China

Bill Wang Biostatistics and Research Decision Sciences, Merck Research Laboratories, Merck & Co., Inc, Kenilworth, NJ, USA

J. Kyle Wathen Quantitative Sciences, Janssen Research & Development LLC, Titusville, NJ, USA

Ying Yuan Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Heng Zhou Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Yanhong Zhou Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Chapter 1
A Statistical Approach to Clinical Trial Simulations

Stephan Ogenstad

1.1 Introduction

Drug development is not for the faintheartedd. We have heard repeatedly over the years regarding the process of bringing a new compound to the market, that every delay will add millions of dollars in added expenses and lost revenues. In order to address some of the concerns in this development process is to simulate the potential outcomes of the clinical study. Simulations of clinical trials go by different names, such as clinical trial simulations (CTS), modeling and simulation (M&S), computer-assisted trial design (CATD), model-based drug development (MBDD), and model-informed drug discovery and development (MID_3). CTS is being increasingly viewed as an integral part of clinical development programs and can be used to improve the understanding and decision making at every stage of drug development. These simulations help to develop better insight into the operating characteristic of a specific trial design. CTS provides the ability to test multiple scenarios, predict the potential study outcomes for each scenario and select the most advantageous study design. Hence, before conducting a study, examining various trial designs through computer simulations can help improve the likelihood of a successful study.

In the field of airplane development, already from the beginning of manned flight, there has been a symbiotic relationship between the airplane and simulation in all of its different forms. The role of simulation and flight simulators in airplane development, training and evaluation have evolved significantly over the past 80 years, often in response to technical innovations in both the airplane and ground support systems. In the same spirit, CTS ought to be an integral part of clinical development, in the design of the clinical protocol and in training of staff and investigators.

S. Ogenstad (✉)
Statogen Consulting LLC, 1600 Woodfield Creek Drive, #215, Wake Forest, NC 27587, USA
e-mail: sogenstad@statogen.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_1

In clinical drug development, the process of development is classified into the four Phases I to IV. Phase I studies are frequently conducted in normal healthy subjects (except for the field of cancer where it is usually done in patients), where focus is on identifying tolerable doses, and on learning about what the body does to the drug (pharmacokinetics) and what the drug does to the body (pharmacodynamics), as well as examining if there are potential interactions with other classes of drugs. In Phase IIA the main objective is to evaluate whether or not the drug has initial encouraging efficacy in a small group of patients ('*proof of principle*' or '*proof of concept*'). The goal of Phase IIB is to learn how to use the drug in a larger group of patients for the indication under consideration. This is usually achieved by applying dose ranging, with or without simultaneous measurements of systemic exposure. In Phase III the efficacy and safety of the novel drug should be confirmed against an established treatment. Sometimes Phase IIIB outcome studies are conducted to learn if, for instance, a type-2 diabetes medication has cardiovascular benefits over other type-2 diabetes medications already on the market. In Phase IV the purpose is to accumulate more information on safety and efficacy from several thousands of volunteers who have the disease. Sheiner has viewed clinical development as two major learn-confirm cycles, the Phase I-IIA and the Phase IIB-III cycles (Sheiner 1997; Sheiner and Ludden 1992).

Nevertheless, even if the main objective of a clinical study is confirming, there are several opportunities to learn about variation in pharmacokinetics and pharmacodynamics in patient groups to increase the likelihood of identifying dosing strategies that will result in safe and effective treatment for the individual patient. Clinical trial simulations can be a valuable tool for decision making in drug development by applying diverse types of models. It then consists of three main components: a disease-placebo model, a drug model, and a trial design model. The disease-placebo model is concerned with the time course of the disease, relative risks with respect to morbidities and mortality. The drug model describes the relationship between therapeutic efficacy, toxicities, and doses. The clinical trial design model deals with components such as baseline characteristics (e.g. inclusion/exclusion, actual values the subjects have at baseline), compliance, missing values, endpoints, and statistical methods of analysis. The use of CTS for drug development has been shown to be a cost-effective approach, for instance, the exploration of multiple dosing regimens and their likely pharmacodynamic effects over diverse patient populations (Huang and Li 2007; Ette et al. 2003; Riggs et al. 2007; Holford and Ploeger 2010). Here, simulations provide a means to assess the effects of various loading and maintenance dosing parameters on steady-state concentrations; effects of dosing holidays (period when a patient is not taking the drug) on pharmacodynamics response; etc.

Without thorough planning, pretesting, and execution, the clinical trial implementation risks are high. Thus, optimization of the clinical trial design should be the main focus before starting the study. In the past, clinical trials were designed using ad hoc empirical approaches, where the '*organization*' impatiently desired the clinical trial to commence under the pretense not to lose any valuable time. Because data resulting from the clinical trial is often too complex to allow simple conclusions of what the outcome of the study is, the interest in CTS has been ongoing for the

past two decades (although there have been earlier success stories recounting the value of simulation for design of clinical trials), and has today become a frequently used tool in quantitative pharmacology investigations in academia, regulatory and the biopharmaceutical industry. Current trends within the pharmaceutical industry and within the offices of some regulatory agencies have suggested a reassuring future for clinical trial simulations (Chang 2010, 2014; Kimko and Peck 2010; Westfall et al. 2008; Duffull and Kimko 2002; Holford et al. 2000; Sheiner and Steimer 2000). If CTS is done thoughtfully, Peck et al. (2003) outline an ambitious but possible future that CTS might sometimes replace the second Phase III trial, and therefore only a single trial is needed.

CTS is the generation of biomarker or clinical responses in virtual subjects that take into account (a) the trial design and execution, (b) pathophysiological changes in subjects during the trial (disease-progress model), and (c) pharmacology (drug-intervention model), using mathematical, statistical and numerical methods and models. CTS can be applied in the design, analysis, and interpretation of human clinical drug trials in order to promote key decisions in drug development management and regulatory approval (Kimko and Peck 2010; Holford et al. 2000). The European Medicines Agency (EMA) and the Center for Drug Evaluation and Research (CDER) in the U.S. Food and Drug Administration (FDA) have each issued a number of guidances for drug developers that pertain to the role of CTS in development and regulation. The FDA's 2009 Guidance for Industry: End-of-Phase 2A Meetings urges sponsors to seek regulatory meetings to discuss quantitative modeling and trial simulations to improve dose selection and clinical trial design. Although not solely focused on CTS, these guidances describe standards and expectations concerning regulatory submission (Kimko and Peck 2010). CTS is included in the FDA's published strategic priorities and is expected to be incorporated in the 2017 PDUFA reauthorization.

Hence, CTS supports the project team to minimize risks and guide decision making by formalizing assumptions, quantifying and testing uncertainties. The simulations can be used for defining and testing analysis models, exploring study design properties, and performing analyses about precision and accuracy of potential endpoint estimates. The simulations can incorporate available scientific information to help the entire project team communicate and test ideas, and to plan significant, effective trials for every phase of clinical development. The CTS helps the team anticipate risks and preview the range of expected results before huge investments are allocated. Thus, CTS has the ability to transform drug development by making better use of prior data and information and to explore important clinical trial designs. As a result, the project team can receive swift feedback on the impact on trial outcomes that alternative designs and analysis methods could have presented in the future. CTS can gain credibility with the '*nonscientists*' as the trial design can be made understandable without technical terms and a different kind of reasoning, and can give clearness to otherwise difficult principles influencing opinion and behavior. The statistician has an imperative role to play within their organization and that by using professionally developed trial design software, such as EAST (Cytel Corporation), or if the organization has invested in the writing of their own computer

programs in, for instance, SAS (SAS Institute) or R. With the help of such software they can rapidly generate many alternative design scenarios that accurately address the questions at hand and the goals of the project team, freeing up time for vital discussions about the choice of endpoints, populations, and treatment regimens.

1.2 Protocol Deviations

Before undertaking any clinical research project, a fully developed and vetted study protocol is critical. In the field of clinical development, having a well written and thought out protocol means that we have a detailed plan that is available and consulted frequently during the conduct of the clinical research project and that the investigators and staff are well trained on at following the protocol. Before the clinical trial starts, it is critical that an efficient statistical methodology is selected and implemented in order to effectively analyze the data after database lock where no data is any longer allowed to be altered. The statistician is critical in conceptualizing the analytical methodology that should be used. Ideally, the statistician needs in a blinded fashion to continue to follow the study as the data is being collected and prior to final analysis of the data. It is not uncommon that the data that was planned to be collected, changes for pragmatic and to some unforeseen reasons. This means that the thoughts that go into the statistical analysis plan should if possible have considered the prospect of such changes could become a reality. Protocol deviations should be rare or unexpected if an intense effort has gone into writing the protocol, though unfortunately many times amendments need to modify the protocols. Consequences of protocol deviations on clinical trial outcomes depend on their qualitative and quantitative characteristics. Thus, while the consequence of one type of protocol deviation can be easily evaluated, some are more difficult to discern than others (e.g. noncompliance to treatment). It follows that the combination of several deviations of varying degrees may lead to unexpected consequences on study outcomes. Protocol deviations can result from many different circumstances, where the most critical deviations are noncompliance and missing data and dropped out subjects.

1.2.1 Noncompliance

Noncompliance or non-adherence to treatment protocol occurs when a patient does not carry out the clinical recommendations of a treating physician. In other words, it is the failure of the patient to follow the prescribed treatment regimen and procedures. Important questions are: What are the consequences if patients take fewer or extra doses of treatment medication than prescribed, but the remaining doses are taken on time, or if patients stop taking the treatment but remain on the study? Noncompliance is a significant problem in all patient populations, from children to the elderly. It applies to nearly all chronic disease states and settings and tends to worsen the

longer a patient continues on drug therapy (Spagnoli et al. 1989; Mardonde et al. 1989; Lacombe et al. 1996). Noncompliance rates with schizophrenia treatment could be as high as 40%, with partial noncompliance as high as 75% (Moore et al. 2000).

Noncompliance can result from a denial of the problem. Many diseases and conditions are easy to ignore, even when they have been diagnosed. This is particularly true for diseases that are asymptomatic and so does not bother the patient. For instance, patients with diabetes, or hypertension may not have symptoms that get in the way of everyday life. They may not even have known that they had the condition until it showed up on a routine examination, which can make it easy for patients to ignore the prescribed treatment regimens. The patients may have difficulty with the regimen and may have trouble following the directions. For instance, taking a pill in the middle of the night, or simply opening the '*child safe*' container may create a barrier to compliance for a patient with rheumatoid arthritis.

Bothersome previous experiences with medications prescribed by their physicians may lead the patients not to take their medication. As a consequence, some patients may not take the medication or may take another medication that they have at home for the same diagnosis. Whether the patients tell the investigators or not will cause difficulties interpreting the results and will bias the study results. Reasons for not disclosing to the investigator that the patient is not taking the medication could be that the patient does not want to affect their relationship with the investigator.

1.2.2 Dropouts and Missing Data

A common problem in clinical trials is the missing data that occurs when patients do not complete the study and drop out without further measurements are taken. Possible reasons for patients dropping out of the study could include death, adverse reactions, unpleasant study procedures, lack of improvement, early recovery, and other factors related or unrelated to trial procedure and treatments. Clinical trials that require adherence that is difficult to follow or have an extensive number of endpoints often suffer from missing data or even subject dropouts. The dropout and missing data mechanisms are often complex, and generally, cannot be assumed to be missing at random or missing completely at random (MCAR). More realistically, the missing values depend on patient experience in the trial. In some cases patient dropouts are infrequent with MCAR mechanism; in other cases, dropouts may be related to a lack of safety or efficacy of the patient's experience. There are several possible ways to model the dropout mechanism; some examples and further references are contained in O'Brien et al. (2005). Patient dropout is a real concern for clinical trials and one of the most problematic protocol deviations. Two types of dropouts exist, non-informative and informative dropouts. Non-informative dropouts simply mean that some patients may randomly stop to be reported in the study, this independently from the treatment they received, and thus independently of efficacy or side effects. Non-informative dropout will simply decrease the statistical study power which is easier to control. On the contrary, disease progress can be perceived by the patient in many

ways not measured in the study but, however, correlate with the endpoint that is being followed. In this case, the dropout is informative to the disease progress, and modeling the disease progress separately from the dropout process may be inefficient and may even produce biased estimates. The bias can be particularly notable if one wants to use the model to predict actually observed features, e.g., observed average disease progress. Imputing unobserved data, e.g., last value carried forward is commonly used as a conservative approach to demonstrate treatment differences, though last value carried forward is, however, inferior from a modeling standpoint as the pseudo data are treated as observed data, creating biases (Westfall et al. 2008).

1.3 Methods

Clinical trial simulations can produce a number of advantages that will help us predict likely outcomes for a range of assumptions about trial size, dose selection and operational considerations, such as:***Study specific aspects***

Study specific aspects

- Comparisons of different trial designs where we can evaluate what we might be losing in one aspect of one design in return for gaining another aspect with another design.
- Optimal dosing for each treatment arm to minimize overlap in exposures and subsequent responses.
- Anticipated patient exposures and responses for each treatment.

Improved specification of inclusion/exclusion criteria

- Optimizing inclusion/exclusion criteria to capture the desired subject population that is influencing the response.
- Potential effects of changes in recruitment rates and criteria on study timelines and results.

Safety and efficacy

- Effects of protocol deviations and treatment compliance on safety and efficacy.

Study results

- Placebo effects on patients over time.
- How the investigational treatment compares to the competitors' treatments.

Statistical analysis

- Whether planned study analysis can detect statistical significance.
- Since conventional statistical tests may be insensitive to a wide range of situations occurring commonly in practice, particularly when the effect of the factor under study is heterogeneous, an evaluation of the test can be made where approximations of the test statistic's distribution have been used in the past.

The computer models that simulate real scenarios are generally developed from previous datasets that may include preclinical data, as well as previous phases of real trials. As clearly stated in Burman et al. (2005) the CTS methodology can be summarized in four steps:

1. Utilizing relevant information.
2. Building a mathematical model (usually for the effect of a drug or device).
3. Predicting the outcome of potential clinical trials.
4. Optimizing the clinical trial program.

Before applying the four steps, the aims of the modeling effort must be defined. What is relevant information, what is a good statistical model, and what is an optimal clinical program depends on the aims we have with the model. The modeling is an interactive process between the formulation of the inputs to the model and the actual outcomes from the simulations. The models should include terms for covariate effects, as models used for simulation studies must deal with the variability from individual to individual. Covariate distribution models describe the relevant information that goes into the simulation, on the basis of preceding trials or clinical experience. The variability of patients' demographic and physiological characteristics in the population of interest that might affect the response. Data in clinical trials are naturally correlated and this should be considered. A number of things about the correlation structures can be learned from previous clinical trials. Baseline measurements are typically correlated with the response. Incorporation of them in the analysis will therefore often considerably improve the trial's effectiveness to show potential therapeutic effects. A baseline response model can help to select the target population or to interpret the trial data. Increasing the number of repeated measurements at baseline and at the end of the treatment period for each subject in a clinical trial will obviously increase the available information on treatment effects and could increase the statistical efficiency of the analysis (Frison and Pocock 1992; Ogenstad 1997). The most efficient way to allocate visits over time at the design stage (e.g., before or after randomization), and the best way to utilize the additional measurements from these visits at the analysis stage is not evident but could be explored via simulations. A model of the baseline response and the variability in the measurements can predict how much the gain would be in terms of efficiency, and could for instance influence the decision on whether the inclusion/exclusion criteria should be modified or not. The impact of the different covariate distributions on the expected outcome of a simulated trial can be assessed, which makes it possible to explore conditions that have been ruled out in the inclusion/exclusion procedures of the actual trial.

As Burman et al. (2005) point out, what information is relevant for the CTS largely depends on what the aims of the modeling are. It also depends on how much information is already available. The best information is perhaps hard endpoint data for the drug in question from a large, randomized, placebo-controlled clinical trial. Unfortunately, this kind of data is seldom available before the end of the clinical program, at the earliest. Hence, what we are concerned with is combining information from diverse sources and incorporating expert judgment in a nonbinding way, and remembering that not all experts are right all the time.

The goal of model building is to establish a model that is fit for the purpose, and not made too involved in order to fulfill the purpose of the design of the clinical study. We need to unify the thinking about the study design and inference. The CTS should make the design and future conduct of the study easier to understand. When a clinical trial is planned, it is supposed that the trial will be executed according to a specific protocol that defines all aspects of the study design, from its beginning to its completion. For instance, characteristics that should be precisely defined in any clinical protocol are whether the subjects are patients or healthy volunteers, inclusion/exclusion criteria, number of subjects to be accrued, treatments and allocation mechanism, blinding of investigators or subjects to the allocated treatment, dosage regimen (dose and timing of doses), endpoints, frequency of follow-up evaluations, and the length of the study.

Complete adherence to the study protocol will permit unbiased estimation of the treatment effects in terms of safety and efficacy with adequate statistical power if the assumptions at the planning stage were correct. Deviations from the protocol may lead to failure of the study to attain its declared purposes. It can be difficult at the planning stage to evaluate what the consequences are of a single protocol deviation, and almost impossible to do it for a combination of protocol deviations. One way to quantify the consequences of those deviations is by using models, describing individual behaviors and responses, combined with trial simulations that include these protocol deviations. When the results of the trial can be envisaged it is sometimes possible to choose, in a methodical and cogent way, between different possible trial designs. The features that are included in the model will unveil what design features can be compared using that model. Missing data cause the usual statistical analysis of complete or all available data to be subject to bias and will diminish the power of the study. Although there are a number of imputation methods, there are no universally applicable methods for handling missing data that will restore the dataset to what it could have been if no data had been missing. As has been noted in the ICH-E9 guideline, '*no universally applicable methods of handling missing values can be recommended*'. The issue of managing missing data is intrinsically difficult because it requires a large proportion of missing data to investigate a method. Moreover, a large proportion of absent data would make a clinical study less credible. The best suggestion is to minimize the chance of dropouts at the design stage and during trial monitoring. It should be reiterated that although an increase in a number of patients to the study will decrease the standard errors, but will not correct the bias that could have been caused due to the missingness of data.

Examples of other features that can be compared are study designs (e.g. sequential, adaptive, crossover, parallel), doses, dosing schedules, study duration, different endpoints, multiple endpoints and timing when the endpoints are measured. The data that is generated from the simulations, together with different statistical methods of analysis of the data may lead to an optimization of the whole study process.

When the purpose of the simulation is to estimate the powers of the statistical tests by the relative number of statistical significances it produces, it is important to use an adequate number of simulations. With 1000 simulations and a power around 90%, the estimation error is approximately 3.7% using a 95% confidence interval. Due to the propagation of uncertainty in the square of a quotient, the uncertainty in the

power estimates translates to an uncertainty in the relative efficiency of the tests in the order of 13–16%. With 10,000 simulations the estimation error is approximately 1.2% and the uncertainty in the relative efficiency of the order of 5%. In pursuance of getting reliable estimates of the true significance level, we recommend simulation sizes around 25,000.

1.4 The Clinical Trial Simulation System

We argue that the CTS system should be flexible, preserving the realism of the doubly-multivariate endpoint/timepoint correlation structures, the informative dropout mechanisms, non-normal distributions, non-monotonic hazard rate functions, survival endpoints, and noncompliance effects (Westfall et al. 2008). The assumptions should be a trade-off between ease of use of the system and realism and flexibility of its outputs. This type of framework for multivariate simulation is usually reasonably simple to program where a variety of software can be used, e.g. SAS (SAS Institute), R (R Foundation for Statistical Computing), SPLUS (TIBCO Software Inc.), Mathematica (Wolfram Research), and MATLAB (MathWorks, Inc.).

From literature, the goal with CTS is often to build a complete model. In Holford et al. (2000) their review on simulations in clinical trials, they state that the model should incorporate all scientific knowledge about the disease and drug. Burman et al. (2005) take a more modest view, where model components should be chosen according to the fit-for-purpose principle. We are convinced that simpler models can sometimes be very useful. Decisions where it may be useful cover a wide range of aspects, including choice of the drug candidate, stop/go for the further development of a compound, choice of patient population, and decisions regarding the positioning versus marketed competitor compounds.

In the PK/PD-phase of drug development, the introduction of population modeling has made it possible through the application of statistical non-linear mixed-effects models to data obtained from relatively few samples in many individuals to discern a genuine insight into the mechanistic aspects (Sheiner and Ludden 1992). More specifically, population models allow characterization of (a) mean pharmacokinetic/pharmacodynamic parameters, (b) extent of variability in these parameters and the sources thereof (e.g. gender, age, disease, comedication), and (c) relationships between pharmacokinetic (e.g. exposure) or pharmacodynamic (e.g. a biomarker) variables and clinical efficacy and safety endpoints. These models can then be used to simulate the outcomes of various trial designs under different assumptions. The usefulness of modeling and simulation in the PK/PD-phase of drug development and regulatory decision-making has been recognized (Holford 1990; Sheiner and Steimer 2000; Holford et al. 2000; Nestorov et al. 2001; Gobburu and Marroum 2001; Gobburu and Sekar 2002; Bhattaram et al. 2005; Burman et al. 2005). Exposure-response models may, for example, be used to support the use of a drug in new target populations through bridging, dose adjustment or no need for dose adjustment in subpopulations, new dose regimens, dosage forms and formulations, routes of administration,

and minor product changes (FDA Guidance for Industry 2003). A biological marker (*biomarker*) has been defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Biomarkers Definition Working Group 2001). The most reliable way to assess the benefit and risk of a drug therapy is through its effect on well-defined clinical endpoints. However, this approach is sometimes impractical for the evaluation of long-term disease therapies and trials that require a large number of patients. A biomarker may then be substituted for clinical response, provided that it is reasonably likely to predict clinical benefit (FDA 1997). However, the single most important use of biomarkers is the selection of the dose range and doses for further investigation in the pivotal trials (Jadhav et al. 2004). To further facilitate the identification of optimal dosing regimens, the use of clinical utility functions has been proposed (Sheiner and Melmon 1978; Eriksen and Keller 1993; Graham et al. 2002; Jonsson and Karlsson 2005). Such functions serve to evaluate important desirable and undesirable effects of a drug on the same scale, under different assumptions of the relative severity of each outcome. In this way, the observed or predicted clinical outcome of different drug therapies, or different dosing regimens of the same drug, may be compared.

An appealing approach to building a statistical CTS system is found in (Westfall et al. 2008). Their approach starts with a model with a rich probabilistic structure to account for typical scenarios, using historical data where it is possible to validate the inputs and outputs, with specific emphasis on the economical yet flexible input of correlation structures. Here, patient responses are functions of underlying correlated $N(0, 1)$ clinical quantities; all distributional forms and dropout effects are determined from these underlying values. Evaluation of trial success then follows from the analysis of the simulated datasets. The goal is to generate realistic datasets having typical correlation structures for multiple endpoint/timepoint data with, say p, endpoints (safety, efficacy or both) indexed by $j = 1, \ldots, p$, and $T + 1$ timepoints indexed by $t = 0, \ldots, T$, where $t = 0$ can be the time of randomization of the patient. For patient i a $p(T + 1)$-vector of correlated $N(0, 1)$ variates Z_{ijt}, each of which may be thought of as a latent indicator of the patient's health relative to a population of similar patients, for endpoint j and timepoint t. Observations will be considered to be independent for different patients. Though, it is possible to include correlations, for instance, for random center effects. Obviously, for each specific patient the timepoint data $Z_{ijo}, \ldots, Z_{ijT}$ are correlated. For instance, the compound symmetry covariance structure model can be expanded easily to accommodate time-series carryover effects in addition to patient effects as $Z_{ijt} = \sqrt{\theta}S + \sqrt{1-\theta}\varepsilon_{ijt}$, where $S \sim N(0, 1)$ is the patient effect and $\varepsilon_{ij0}, \ldots, \varepsilon_{ijT}$ is a realization of a unit variance AR(l) process with parameter p. For simulation purposes, the parameters θ and p must be specified. For multiple endpoint data for patient and timepoint, it is suggested that the correlation between endpoints is best left as unstructured. For each patient, the observations between endpoints at different timepoints are correlated. There are a number of possibilities for defining this structure, the most convenient and well-known is the Kronecker product model used in multivariate longitudinal models (Westfall et al. 2008).

Most commercially available clinical trial software systems use parametric input into the systems. For instance, the exponential survival model is often used as input model. Though, the exponential survival model is a rather unrealistic model since it is assumed that the hazard rate function is constant over the entire observational study period. The Weibull model is many times a better choice than the exponential, but this model still has a monotonic hazard rate function, which might not be realistic either. A more flexible approach is to use Royston-Parmar models (Royston and Parmar 2002) that have great flexibility. Even better at times is to use mean structures as input for the different endpoint*timepoint*treatment combinations. Such structures can be determined purely a priori from earlier phase data, suggested by PK/PD models, or from studies on similar interventions. Survival analyses pose additional questions. Standard methods such as the log-rank test and Cox models are efficient when the hazards are proportional. This assumption is not always reasonable. The non-proportional hazards assumption that is a potential difficulty with the Cox model, could sometimes be handled in a simpler way, and the visualization of the hazard rate function could be made easier, using the Royston–Parmar framework. In Westfall et al. (2008), any types of distributions could be applied to the mean structures, and there they effectively made use of a missing value, dropout and noncompliance mechanism to generate '*real world datasets*'. Girard et al. (1998) developed a hierarchical Markov model for patient compliance with oral medications conditional upon a set of individual-specific nominal daily dose times and individual random effects that are assumed to be multivariate normally distributed. This model also has great flexibility and allows descriptions of almost all possible compliance profiles.

1.5 Some Published Clinical Trial Simulations

Wathen and Thall (2008) presented a new approach to the problem of deriving an optimal design for a randomized group sequential clinical trial based on right-censored event times. They were motivated by the fact that, if the proportional hazards assumption is not met, then a conventional design's actual power can differ substantially from its nominal value, and combined Bayesian decision theory, Bayesian model selection, and simulation to obtain a group sequential procedure that maintains targeted false-positive rate and power, under a wide range of true event time distributions. At each interim analysis, the method adaptively chooses the most likely model and then applies the decision bounds that are optimal under the chosen model. A simulation study comparing this design with three conventional designs showed that, over a wide range of distributions, their proposed methods perform at least as well as each conventional designs, and in many cases, it provides a much smaller trial.

Dragalin et al. (2010) presented a simulation study to compare new adaptive dose-ranging design. The main goals in an adaptive dose-ranging study are to detect dose-response, to determine if any doses meet clinical relevance, to estimate the dose-response, and then to decide on the dose(s) (if any) to take into the confirmatory Phase III. Adaptive dose-ranging study designs may result in power gains to detect dose-

response and higher precision in estimating the target dose and the dose response curve.

Kimko et al. (2000) simulated the anticipated results of a Phase III clinical trial of the antischizophrenic drug, quetiapine, based on input-output and covariate distribution models developed using data collected in earlier Phase I and II trials. The model development was performed using the NONMEM program with first order conditional estimation (Beal and Sheiner 1992). The proposed trial design was a double-blind, placebo-controlled, randomized, parallel group study of fixed-dose of quetiapine in hospitalized schizophrenic patients, who received one of five doses of quetiapine or placebo for a period of four weeks. The treatment was initiated after a placebo run-in period followed by a two week step-wise dose titration period. The executed study design was replicated by excluding individuals wrongly included in the study, as they failed to meet the entry criteria. In addition, placebo responders identified during the placebo run-in period were replaced. A random dropout algorithm using a multiplicative congruential method (such that the random number generated is the remainder of a linear transformation of the previous number divided by an integer) was used to simulate the high dropout rate observed in the earlier Phase II study. Based on the Phase II study result, 70% of the patients assigned to the placebo group, 60% assigned to the lowest dose group and 50% assigned to all other dose groups were withdrawn from the study. Simulations were performed for 100 sets of 50 patients per treatment group. Adequacy of the model to describe the original data was tested using sensitivity analysis and by comparing posterior parameter distributions and posterior predictions from the simulated trial design to parameters of the prior distribution and observed data. Dropout rates in the simulation and in the Phase III trial were comparable. Comparison of the simulated results with actual results obtained in the Phase III trial showed that the model adequately predicted responses to quetiapine. However, it was found to be inadequate in predicting the placebo response.

Clinical trial simulation for docetaxel was performed using pharmacokinetic/pharmacodynamic models previously developed from data obtained in earlier open-label, non-randomized, Phase II clinical trials of docetaxel in subjects with small cell lung cancer. The purpose of the simulation was to predict the influence of dose on survival time and time to disease progression in a high-risk group in a planned Phase III trial comparing doses of docetaxel of 100–125 mg/m^2 every three weeks. Input-output and covariate distribution models were developed using the NONMEM program. Hazard models were used to simulate the primary and secondary clinical endpoints, death and disease progression, respectively. In addition, the execution model included a separate hazard model for patient dropout. Different models were tested and the Weibull distribution was selected based on the goodness of fit assessed in the model-building phase of the analysis. A dose titration algorithm allowed for a 25% dosage reduction in the event of severe toxicity for each treatment cycle. To maintain consistency with study implementation, after two dosage reductions or if disease progression occurred, the patient was withdrawn from the study. Simulations were performed for 100 sets of subjects and the results were analyzed using SAS. Adequacy of the model to describe the Phase II data was tested using a posterior

predictive check of the following test quantities: number of deaths and progressions, median survival time, I year survival, median time to progression, patient characteristics at baseline, number of side-effects at the end of the first cycle, number of treatment cycles per patient and total dose. Tabulated median and 95% confidence intervals of simulated test quantities agreed well with those obtained from the original data. In addition, 100 sets of 200 subjects per treatment group were simulated under the Phase III trial design and test quantities were calculated. The results of the Phase III trial simulation showed no clinical advantage of the higher docetaxel dose on survival or time to disease progression in high-risk subjects with small cell lung cancer. As a consequence of this analysis, it was determined that there would be no further clinical studies to evaluate the effect of dose intensification in subjects with small cell lung cancer.

1.6 Commercially Available Trial Design Software Packages

Performing simulations with most currently available simulation tools is an investment of time, requiring custom programming and at times moving between one software application to perform simulations and another application to visualize simulations. There is a great need for even more efficient simulation systems that facilitate interactive, real-time evaluation and iteration on simulation scenarios.

As indicated earlier, more adaptations give the investigator more flexibility in identifying best clinical benefits of the test treatment under investigation. However, multiple adaptive designs with more adaptations could be very complicated and consequently, appropriate statistical methods for assessment of the treatment effect may not be available and are difficult, if not impossible, to obtain. Thus, one of the major obstacles for implementing adaptive design methods in clinical trials is that the appropriate statistical methods are not well established with respect to various adaptations. Though, some practical methods in this field are emerging (Gao et al. 2013; Pong et al. 2010). Current software packages such as SAS cannot be applied directly and hence are not helpful here. Although there are some software available in the marketplace such as ExpDesign Studio (http://www.ctrisoft.net), EastSurvAdapt (Cytel Corporation), and ADDPLAN (http://www.addplan.com), which cover certain types of adaptive trial designs, new software packages for adaptive design methods in clinical trials are necessary to assist in implementing adaptive trial designs in clinical trials (Wassmer and Vandemeulebroecke 2006). An overview of software available for group sequential and adaptive designs can also be found in Herson (2009).

Some software (e.g., Certara, https://www.certara.com/software/; Lixoft, http://lixoft.com/) require PK/PD input as drivers for the simulation output. A well developed system is found with EAST 6 from Cytel Corporation that has a large variety of parametric design choices. Another software that produces data with '*flexible*'

statistical characteristics, which helps the decision making that statisticians typically must make is developed by Westfall et al. (2010).

Concerning the description of virtual patients, i.e. the distribution of covariates in a target population, general-purpose statistical packages can be employed. Note that, since IO models usually include terms for covariate effects, the choice of methodology for generating virtual subjects is often dependent on the software for IO modeling. Mouksassi et al. (2009) use the R package library GAMLSS, which facilitates the simulation of demographic covariates specific to the targeted patient populations. Other authors (Chabaud et al. 2002) prefer to resample patients from existing epidemiological databases rather than creating realistic virtual subjects.

The R software environment (by R Core Team 2014) has an excellent set of tools for analyzing and visualizing simulation results in real time. The new RxODE package facilitates quick and efficient simulations of ordinary differential equation (ODE) models in R. RxODE provides an elegant, efficient, and versatile way to specify dosing scenarios, including multiple routes of administrations within a single regimen, sampling schedules, etc. It also enables simulations with between-patient variability and minimizes the amount of custom coding required for pharmacometrics simulations (Wang et al. 2015).

A system specifically designed for IO-modeling of data in this context are the non-linear mixed-effect model program NONMEM (developed by Stuart L. Beal and Lewis B. Sheiner in the late 1970s at UCSF for population pharmacokinetic modeling). It is still widely used.

ADAPT (Biomedical Simulations Resource (BMSR) in the Department of Biomedical Engineering at the University of Southern California) is a computational modeling platform developed for PK/PD applications. It is intended for basic and clinical research scientists and is designed to facilitate the discovery, exploration and application of the underlying pharmacokinetic and pharmacodynamic properties of drugs, which includes an extensive library of models to choose from.

MATLAB (MathWorks) is a multi-paradigm numerical computing environment and fourth-generation programming language. MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C#, Java, Fortran and Python. MATLAB provides a software tool, the so-called SimBiology, for the complete PK/PD workflow. Since Sim-Biology is based on MATLAB, users can employ MATLAB in order to program their simulations.

Mathematica (Wolfram Research) is a quality symbolic computation system. For clinical trial simulations SystemModeler is excellent for modeling and analysis throughout drug discovery, development, clinical trials, and manufacturing. The flexible environment supports application areas such as systems biology, bioinformatics, and more.

Mathematica and MATLAB are very different products. Mathematica focuses on quality symbolic computation and features like unlimited precision arithmetic. MATLAB focuses on high speed algorithms for numerical computation.

1.7 Discussion

In the past few years, scientific journals covering clinical pharmacology and pharmacokinetics and trials in later phases have published a large number of papers related to CTS. The interest in CTS in statistical literature within statistical university departments has been much lower. Still, statistics and statisticians are needed in CTS activities. By writing this paper, we would like to stimulate more statisticians to take an active part in applied modeling work and research related to CTS.

Some of the examples given have hopefully shown that even quite simple modeling exercises can prove very useful. One task for the modeler is precisely that of finding those questions where a limited amount of work is likely to give significant benefits. It might be hard for some statistical scientists to accept that being too rigorous may be harmful. The model need not be perfect. What matters is that the work is good enough to help make the right decisions.

Even though practical modeling work may sometimes be "quick and dirty", rigorous statistical research is needed in the CTS area. We would especially like to point out the need to apply and integrate different areas within statistics and to integrate statistical results into other disciplines, such as pharmacometrics and pharmacoeconomics.

CTS integrates expert knowledge in the relevant fields (primarily pharmacology and medicine in the clinical phase) with new data in a structured process to create quantitative models. The cooperation between different skills is thus essential. Some modeling work can, of course, be done by a single individual. In many situations, however, the greatest benefits are likely to result from a joint collaboration with several skills working in concert (e.g., Biomarkers Definition Working Group). What skills to include in the modeling team is, of course, depending on the modeling questions. Good organization is critical both internally in the modeling team and for the team's relations with decision makers and experts from different parts of the research organization.

CTS aims at optimizing a clinical development program. This program, however, is not totally isolated from the rest of drug development and commercialization. What is 'optimal' in clinical development depends on factors such as the medical need for a new treatment, its commercial value, the regulatory requirements, and the ability to find patients and produce tablets in time for the clinical trials etc. CTS should therefore not be seen as separate from other modeling activities. Preclinical, epidemiological and commercial models could provide useful input to CTS. The results of CTS, on the other hand, may be of great value for predicting market penetration and sales.

Execution models are used to examine the influences of protocol deviations on study outcomes. When implemented as a part of a clinical trial simulation, they allow "virtual" clinical trials to the run under varying conditions, from simple errors in data gathering to complex combinations of protocol deviations that emulate real-world situations. Thus, execution models arc powerful tools for identifying weaknesses or limitations in a proposed study design, which may be anticipated, avoided or resolved

in order to increase the robustness of the study design prior to implementation of the actual clinical study. As such, they are an integral component of clinical trial simulation and essential tools for identifying weaknesses or limitations in a proposed study design, which may be anticipated, avoided or resolved in order to increase the robustness of the study design prior to implementation of the actual clinical study. As such, they are an integral component of clinical trial simulation and an essential tool in clinical trial design. Execution models for protocol departures do not necessarily require data to be identified, except for dropout. Many trials can be performed in simulators that are just too risky in real life and they can be repeated multiple times. Simulators tend to prevent trial failures or overpowered studies by their ability to point what part of the experiment is the most sensitive to protocol departures. Indeed clinical trial simulation provides an invaluable tool to prospectively force experimental study designs to the point of failure.

References

Beal, S. L., & Sheiner, L. B. (1992). *NONMEM user's guide*. San Francisco: NONMEM Project Group, University of California, San Francisco.

Bhattaram, V. A., Booth, B. P., Ramchandani, R. P., Beasley, B. N., Wang, Y., Tandon, V., et al. (2005). Impact of pharmacometrics on drug approval and labeling decisions: a survey of 42 new drug applications. *American Association of Pharmaceutical Scientists Journal, 7,* E503–E5122005.

Biomarkers Definition Working Group. (2001). Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics, 69*(3), 89–95.

Burman, C. F., Hamren, B., & Olsson, P. (2005). Modelling and simulation to improve decision-making in clinical development. *Pharmaceutical Statistics, 4*(1), 47–58.

Chabaud, S., Girard, P., Nony, P., & Boissel, J. P. (2002). Clinical trial simulation using therapeutic effect modeling: Application to ivabradine efficacy in patients with angina pectoris. *Journal of Pharmacokinetics and Pharmacodynamics, 29*, 339–363.

Chang, M. (2010). *Monte Carlo simulation for the pharmaceutical industry: Concepts, algorithms, and case studies*. Chapman & Hall/CRC Biostatistics Series.

Chang, M. (2014). *Adaptive design theory and implementation using SAS and R*. CRC Press.

Dragalin, V., Bornkamp, B., Bretz, F., Miller, F., Padmanabhan, S. K., Patel, N., et al. (2010). A simulation study to compare new adaptive dose-ranging designs. *Statistics in Biopharmaceutical Research, 2,* 487–512.

Eriksen, S., & Keller, L. R. (1993). A multiattribute-utility-function approach to weighing the risks and benefits of pharmaceutical agents. *Medical Decision Making, 13*(2), 118–125.

Ette, E., Godfrey, C., Ogenstad, S., & Williams, P. (2003). Analysis of simulated clinical trials. In H. Kimko, & S. Duffull (Eds.), *Simulation for designing clinical trials: A pharmacokinetic/pharmacodynamic modeling perspective*. New York: Marcel Dekker.

Food and Drug Administration. (1997). *The Food and Drug Modernization Act of 1997, in Code of Federal Regulations Title 21 Part 314 Subpart H*.

Food and Drug Administration, Guidance for Industry. (2003). *Exposure-response relationships—Study design, data analysis, and regulatory applications* (pp. 1–25). Rockville, MD.

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

Gao, P., Liu, L., & Metha, C. (2013). Exact inference for adaptive group sequential designs. *Statistics in Medicine*.

Gobburu, J. V. S., & Marroum, P. J. (2001). Utilisation of pharmacokinetic/pharmacodynamic modeling and simulation in regulatory decision-making. *Clinical Pharmacokinetics, 40*(12), 883–892.
Gobburu, J. V. S., & Sekar, V. J. (2002). Application of modeling and simulation to integrate clinical pharmacology knowledge across a new drug application. *International Journal of Clinical Pharmacology and Therapeutics, 40*(7), 281–288.
Graham, G., Gupta, S., & Aarons, L. (2002). Determination of an optimal dosage regimen using a bayesian decision analysis of efficacy and adverse effect data. *Journal of Pharmacokinetics and Pharmacodynamics, 29*(1), 67–88.
Herson, J. (2009). *Data and safety monitoring committees in clinical trials*. Boca Raton: Chapman & Hall/CRC.
Holford, N. H. G. (1990). Concepts and usefulness of pharmacokinetic/pharmacodynamic modeling. *Fundamental & Clinical Pharmacology, 4*(Suppl. 2), 93S–101S.
Holford, N., & Ploeger, B. A. (2010). Clinical trial simulation: A review. *Clinical Pharmacology and Therapeutics, 88,* 166–182.
Holford, N. H. G., Kimko, H. C., Monteleone, J. P. R., & Peck, C. C. (2000). Simulation of clinical trials. *Annual Review of Pharmacology and Toxicology, 40,* 209–234.
Huang, X., & Li, J. (2007). Pharmacometrics: The science of quantitative pharmacology. *American Journal of Pharmaceutical Education, 71,* 75.
Jadhav, P. R., Mehta, M. U., & Gobburu, J. V. S. (2004). How biomarkers can improve clinical drug development. *American Pharmaceutical Review, 7*(3), 62–64.
Jonsson, S., & Karlsson, M. O. (2005). Estimation of dosing strategies aiming at maximizing utility or responder probability, using oxybutynin as an example drug. *European Journal of Pharmaceutical Sciences, 25*(1), 123–132.
Kimko, H., & Duffull, S. B. (2002). *Simulation for designing clinical trials: A pharmacokinetic-pharmacodynamic modeling perspective*. CRC Press.
Kimko, H., & Peck, C. C. (2010). *Clinical trial simulations: Applications and trends*. Springer.
Lacombe, P. A., Vicente, J. A. G., Pages, J. C., & Morselli, P. L. (1996). Causes and problems of non response or poor response to drugs. *Drugs, 51,* 552–570.
Maronde, R. F., Chan, L. S., Larsen, F. J., Strandberg, L. R., Laventurier, M. F., & Sullivan, S. R. (1989). Underutilization of antihypertensive drugs and associated hospitalisation. *Medical Care, 27,* 1159–1166.
Moore, A., Sellwood, W., & Stirling, J. (2000). Compliance and psychological reactance in schizophrenia. *British Journal of Clinical Psychology*.
Mouksassi, M. S., Marier, J. F., Cyran, J., & Vinks, A. A. (2009). Clinical trial simulations in pediatric patients using realistic covariates: Application to teduglutide, a glucagon-like peptide-2 analog in neonates and infants with short-bowel syndrome. *Clinical Pharmacology & Therapeutics*.
Nestorov, I., et al. (2001). Modeling and simulation for clinical trial design involving a categorical response: A phase II case study with naratriptan. *Pharmaceutical Research, 18*(8), 1210–1219.
O'Brien, P. C., Zhang, D., & Bailey, K. R. (2005). Semi-parametric and non-parametric methods for clinical trials with incomplete data. *Statistics in Medicine, 24,* 341–358.
Ogenstad, S. (1997). Analysis and design of repeated measures in clinical trials using summary statistics. *Journal of Biopharmaceutical Statistics, 7*(4), 593–604.
Peck, C. C., Rubin, D. B., & Sheiner, L. B. (2003). Hypothesis: A single clinical trial plus causal evidence of effectiveness is sufficient for drug approval. *Clinical Pharmacology & Therapeutics, 73,* 481–490.
Pong, A., & Chow, S. C. (2010). *Handbook of adaptive designs in pharmaceutical and clinical development*. Taylor & Francis, New York: Chapman and Hall/CRC Press.
R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. [Software] (Version 3.1.1). http://www.R-project.org/.
Riggs, M. M., Godfrey, C. J., & Gastonguay, M. R. (2007). Clinical trial simulation: efficacy trial. In E. I. Ette, & P. J. Williams (Eds.), *Pharmacometrics: The science of quantitative pharmacology* (pp. 881–900). Hoboken, NJ: Wiley.

Royston, P., & Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine, 21,* 2175–2197.

Sheiner, L. B. (1997). Learning versus confirming in clinical drug development. *Clinical Pharmacology and Therapeutics, 61*(3), 275–291.

Sheiner, L. B., & Ludden, T. M. (1992). Population pharmacokinetics/dynamics. *Annual Review of Pharmacology and Toxicology, 32,* 185–209.

Sheiner, L., & Melmon, K. (1978). The utility function of antihypertensive therapy. *Annals of the New York Academy of Sciences, 304,* 112–127.

Sheiner, L. B., & Steimer, J.-L. (2000). Pharmacokinetic/pharmacodynamic modeling in drug development. *Annual Review of Pharmacology and Toxicology, 40*(1), 67–95.

Spagnoli, A., Ostino, G., & Borga, A. D. (1989). Drug compliance and unreported drugs in the elderly. *Journal of the American Geriatrics Society, 37,* 619–624.

The International Conference on Harmonization (ICH) http://www.ifpma.org/ich1.html.

Wang, W., Hallow, K. M., & James, D. A. (2015). A tutorial on RxODE: Simulating differential equation pharmacometric models in R. *American Society for Clinical Pharmacology and Therapeutics*. CPT: Pharmacometrics & Systems Pharmacology.

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

Wathen, J. K., & Thall, P. F. (2008). Bayesian adaptive model selection for optimizing group sequential clinical trials. *Statistics in Medicine, 27*(27), 5586–5604.

Westfall, P. H., Tsai, K., Ogenstad, S., Tomoiaga, A., Moseley, S., & Lu, Y. (2008). Clinical trials simulation: A statistical approach. *Journal of Biopharmaceutical Statistics, 18,* 611–630.

Westfall, P. H., Dmitrienko, A., DeSouza, C., & Tomoiaga, A. (2010). *Clinical trials simulation: A publicly available, grid-enabled, GUI-Driven SAS® System*. SAS Global Forum 2010, Paper 180-2010.

Chapter 2
Choosing the Function of Baseline Run-in Data for Use as a Covariate in the Analysis of Treatment Data from Phase III Clinical Trials in Hypertension

Yi Hao and Karl E. Peace

2.1 Introduction

High Blood pressure or hypertension is a common chronic disease among all ages of populations. An individual who has high blood pressure usually has some other concurrent disease, such as diabetes, renal disease, cancer or other cardiovascular disease (Nussbaumerov and Rosolov 2013). Controlling blood pressure in such individuals is a key issue in their clinical management. In clinical trials of hypertension, for example, blood pressure is measured just prior to randomization (baseline) to ensure that the volunteer is eligible for entry into the trial.

Eligible patients are then randomized to the clinical trial treatment groups and followed over the length of the treatment period as specified by the trial protocol. Blood pressure is measured after randomization when patients return to the clinic at scheduled visits. Although both systolic and diastolic blood pressure are measured, diastolic blood pressure (DBP) at post randomization visits or change from baseline is analyzed to assess the antihypertensive efficacy of treatment interventions using baseline diastolic blood pressure as a covariate (Blood Pressure Lowering Treatment Trialists' Collaboration 2003).

In terms of diastolic blood pressure, there are different stages of hypertension. Prehypertension is a diastolic pressure ranging from 80 to 89 mm Hg. Stage 1 hyper-

Y. Hao
Frontier Science and Technology Research, 440 Science Drive suite 401, Madison, WI 53711, USA

K. E. Peace (✉)
Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA 30461, USA
e-mail: Peacekarl@frontier.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_2

tension is a diastolic pressure ranging from 90 to 99 mm Hg. Stage 2 hypertension is a diastolic pressure of 100 mmHg or higher.[1]

There are many trial designs that could be used for different testing purposes. Prospective, randomized, double-blind, placebo controlled design is most commonly used in phase III clinical trials. In addition, adaptive designs and sequential designs are two very popular trial designs in the pharmaceutical industry. In this paper, we use the typical prospective, randomized, double-blind, placebo-controlled design to simulate diastolic blood pressure in previously untreated patients newly diagnosed with prehypertension ($80 < DBP < 90$), stage 1 hypertension ($90 \leq DBP < 100$) or stage 2 hypertension ($100 \leq DBP < 120$), where DBP is measured in mmHG.

A concern in clinical trials of hypertension is that a single baseline value may not adequately reflect the true severity of the disease prior to randomization, and hence may not be the best choice of a covariate. However, a single baseline value (or the average of three consecutive measurements a few minutes apart in hypertension studies) is what usually available (Winkens et al. 2007) is. It is therefore compelling to have measurements of disease severity over a baseline run-in period to better characterize the natural progress of the disease.

Therefore, an open question is: What is the best function of the baseline run-in data to use as a covariate in the assessment of treatment group differences post randomization? In this paper, we assume that the length of the baseline run-in period is 8 days (approximately a week-sufficient to establish the blood pressure profile prior to treatment) and that diastolic blood pressure, the primary measure of hypertension, is measured on each eligible patient on each of the 8 days. We then explore several functions of the baseline run-in data for use as a covariate in the analysis of post randomization data in order to determining which covariate is optimal (maximum power or minimum MSE).

To address the above two questions, we simulate diastolic blood pressure data reflecting patients newly diagnosed with hypertension, followed over 8 consecutive days of baseline run-in (essentially on placebo for 8 days) and then monthly for six months of treatment with either placebo or a drug being developed for the treatment of hypertension. At the end of the baseline run-in period, patients are randomized in balanced fashion to placebo or drug, provided medication for one month and instructed to return to the clinic every month for clinical evaluation and remedication. Both patients and investigational site personnel are assumed to be blinded as to identity of intervention received during the 6 months of treatment. This mimics a Phase III pivotal proof of efficacy trial in hypertension. We assume that the attendant protocol has been approved by an IRB. Here our objective is to determine what function of 8 days of baseline run-in data is best to use as a covariate in the analysis of post-randomization treatment group data.

[1] High blood pressure (hypertension), Mayo Clinic.

2.2 Literature Review

In blood pressure trials, examples of covariates are sex, age and weight (or body surface area), but so is diastolic blood pressure measured at baseline just prior to randomization. Some authors have discussed whether the value of blood pressure at baseline should be used as a covariate or as a dependent variable? Liu et al. (2009) used two statistical methods: the first was to provide a point estimate and a 95% confidence interval for the mean change from baseline at time T for each treatment group; the second was to provide a p-value and a 95% confidence interval for the between-group difference in the mean changes from baseline. In Liu's paper, post-randomization measures were repeated, while baselines were not repeated measures. Assmann et al. (2000), Pocock et al. (2002) reported use of baseline data for (i) subgroup analyses which explore whether there is evidence of a treatment difference; (ii) Covariate-adjusted analyses which aim to refine the analysis of the overall treatment difference with the fact that some baseline characteristics are related to outcome and may be unbalanced between treatment groups; and (iii) baseline comparisons which compare the baseline characteristics of patients in each treatment group for any possible difference. Not all baseline variates may be used directly, especial in non-linear dose-response models. Box-Cox transformation (Chen and Pounds 1998) would be a reasonable method for rendering such variates to a proper form, i.e. log transformation.

Alternatively, if the baseline variate reflects a repeated measures structure, for example, blood pressure trials, it is very common that investigators would collect more than one baseline blood pressure measurement. The length of the baseline period may vary from case to case. However, the last measurement of the last baseline period is typically the first choice as a covariate. David (Bristol 2007) talked about the choice of two baselines (X_1, X_2) as a covariate in performing between-treatment comparisons. He assumed (X_1, X_2, Y) followed a trivariate normal distribution with $EX_1 = EX_2 = 0$, heterogeneous variances and different correlations. His results show that using both baseline measurements as covariates increased the power of the analysis of treatment group difference. Zhang et al. (2010) also talked about the choice of two baseline measurements in his paper. By using ANCOVA, under a simulation study, he assumed there are two baseline measurements and one post-baseline measurement (response in the ANCOVA). Four functions of baseline as covariates were used in this simulation to determine which one gave the largest power. Five models were investigated either by classical ANCOVA or logistic regression. They were no covariate, first baseline as covariate, second baseline as covariate, both baselines as covariates, and average of the baselines as covariate. Simulations were performed using the multivariate normal distribution under specified variance-covariance matrices. He also gave an example to show how to compute means and variance-covariance matrices under specified correlations. He concluded that including both baseline measurements were consistently best in increasing power and reducing mean square error (MSE) in the Covariance analysis of an efficacy response variable.

Other papers discussed using ANCOVA and ANOVA to analyze treatment group differences and testing for treatment efficacy. Many papers discussed the use of covariates in data analyses, but few if any, have addressed the choice of covariates as functions of repeated measurements over a baseline run-in period. Our research goal is to fill this gap by finding the best covariate as a function of the repeated baseline measurements to improve clinical trial testing of drug effectiveness.

2.3 Theory

In clinical trials, variables or data taken at baseline are often used as covariates in statistically analyzing post-randomization treatment group data. Sex, age, weight or other demographic factors such as disease severity on trial subjects are often used as covariates. After identifying the covariates, the analysis of covariance (ANCOVA) is used as the analysis method to test differences between the target treatment and the comparator.

There are two main purposes for covariate adjustment in statistical analyses: (1) To improve the credibility of the trial results by demonstrating that any observed treatment effect is not accounted for by an imbalance in patient characteristics; and (2) To improve statistical efficiency (Tu et al. 2000).

Covariates play a very important role in the design and analysis of clinical trials. In the estimation of a treatment effect in a randomized trial, the role of covariates is to adjust for any bias associated with covariate imbalance at baseline and to improve the precision of the estimated treatment effect, independent of the nature of the model being fit to the data (Ford and Norrie 2002).

In this paper, we will discuss various functions of repeated baseline measurements taken prior to randomization to be used as covariates in the analysis of treatment group differences of post randomization data. When there are more than two baseline measurements there can be multiple functions of them that may be used as covariates. Some functions are: mean, median, maximum, minimum, area under the curve (AUC) and so on. Once the functions of the baseline measurements have been calculated, each individually may be added to the ANOVA model, to produce ANCOVA models. The post randomization data is then analyzed using both the ANOVA and ANCOVA models to assess treatment group differences and mean-square-error compared between the ANOVA and each of the ANCOVA models. The covariate leading to the smallest MSE would be considered the best among all the baseline covariates. We consider a hypertension clinical trial with diastolic blood pressure as the continuous (dependent) response variable. We simulate DBP data on subjects to reflect that of patients who have hypertension. Baseline diastolic blood pressure is measured once a day for 8 consecutive days.

2.4 Patterns of Data

We identify two classes of data patterns over the baseline run-in period. The patterns may be linear or non-linear. For each class, we have three cases, respectively, for a total of 6 cases. We design the clinical trial, baseline, post-baseline, and discuss how to simulate data sets for the 6 cases.

We simulate a clinical trial in hypertension that has three phases: (1) A baseline run- in period of 8 days in length; (2) Then randomization of patients into a drug group and a placebo group; and (3) Then a treatment period of 6 months. We assume that diastolic blood pressure is measured on each patient once on each day of baseline, and then monthly, up to six-months during the treatment period. DBP measured during the baseline run-in period provides data for possible covariates, whereas DBP measured during the treatment period provides response data that may be analyzed for assessing treatment differences.

We describe below the patterns of baseline run-in data and patterns of treatment period data we simulate and analyze to address the objectives of this study.

2.4.1 Baseline Run-in, Pre-randomization Period

In general, data over the baseline run-in period may exhibit linearity or non-linearity over the number of days. Then linearity or non-linearity may reflect different forms. Linearity could reflect constancy, strictly increasing or strictly decreasing over the baseline run-in period. Non-linearity could manifest in an oscillatory manner, or as a mixture of linear and non- linear components or in a random-walk fashion. These cases cover practically all baseline run-in data behavior in clinical trials in hypertension.

2.4.1.1 Linear Cases

For the linear patterns we consider three cases: (1) the mean of the data on patients on each day is constant over the number of days; (2) the mean of the data on patients on each day reflects an upward trend over the number of days; and (3) the mean of the data on patients on each day reflects a downward trend over the number of days.

Case 1: The Data Are Constant Over the Number of Days

When plotting the baseline run-in data by day, although the data per day reflects inter- subject variation, the means of the data across days are constant (slope of least squares line is not significantly different from 0).

Case 2: The Data Reflect an Upward Trend Over the Number of Days

When plotting the baseline run-in data by day, although the data per day reflects inter- subject variation, the means of the data across days reflects an upward trend (slope of least squares line is positive and significantly different from 0).

Case 3: The Data Reflect a Downward Trend Over the Number of Days

When plotting the baseline run-in data by day, although the data per day reflects inter- subject variation, the means of the data across days reflects a downward trend (slope of least squares line is negative and significantly different from 0).

2.4.1.2 Non-linear Cases

We also consider three non-linear cases: (1) the mean of data on each baseline run-in day is periodic as described by the sine function; (2) a mixture of all linear cases and the periodic non-linear case as one data set; and (3) the mean of data on each baseline run-in day acts as a random walk, with no specific trend.

Case 4: The Data Are Periodic as Reflected by the Sine Function

When plotting the baseline run-in data by day, although the data per day reflects inter-subject variation, the means of the data across days reflects periodicity, for example $X = \mathrm{Sin}(D)$, where X is the baseline data and D is the day of the baseline run-in period when measurements are made.

Case 5: The Data Reflect Mixed Linear and Non-linear Data

Based on the data sets generated before reflecting constant, linear increase, linear decrease and oscillatory as per the Sine function, we randomly select 20–25% subjects from these four baseline data sets separately, and combine them into one mixture data set. This is a nonparametric data set with mean and variance computed using the existing simulated data.

Case 6: The Data Derive from a Random Walk

When plotting the baseline run-in data by day, although the data per day reflects inter- subject variation, the means of the data across days have no specific or obvious trend and move as a random walk.

2.4.2 Post Randomization Treatment Period

At the end of the baseline run-in period, for all cases (three linear and three non-linear), we randomize all subjects into a pseudo-placebo group (referred to as Placebo) and a pseudo-treatment group (referred to as Treatment) in balanced fashion (50 per arm). Data are then simulated over a six-month period that mimics subjects in the Treatment group receiving an anti-hypertensive drug and subjects in the Placebo group receiving a matching placebo. There will be 100 post-baselines at each month-for a total of 600 post-baseline DBP measurements over the six months treatment period.

We assume that there is a gradual decrease over the six months treatment period in both the Placebo and Treatment groups. For the Placebo group, we assume mean DBP to be 90 mmHg at the end of the treatment period-an average decrease of 5 mmHg in DBP over the six months treatment period. For the Treatment group, we assume mean DBP to be 80 mmHg at the end of the treatment period-an average decrease of 15 mmHg in DBP over the six months treatment period. Note that in Sect. 2.5.4, the mean of the distribution of DBP just prior to randomization (after measurement on baseline run-in day 8), is above or around 95 mmhg, for the three linear and three non-linear cases considered over the baseline run-in period.

2.5 Simulation

2.5.1 Ranges and Size of the Cohort Over Baseline Run-in

The size of the cohort over the baseline run-in period is chosen to be 100 patients. These patients will be randomized to two groups in balanced fashion at the end of the baseline run-in period and treated for 6 months. Patients in one group will be assumed to receive an antihypertensive drug; whereas patients in the other group will be assumed to receive a placebo. Fifty patients per group will provide a power of approximately 99% to detect 10 mmHg greater reduction in the treatment group than in the placebo group by the end of six months of treatment. So 100 patients participating in the baseline run-in period should provide sufficient power to detect a clinically meaningful difference between an antihypertensive medication and placebo over 6 months of treatment.

We assume that the screening range of DBP on each subject is (80,120]. As previously mentioned this allows patients with pre-hypertension, stage 1 hypertension or stage 2 hypertension to be studied. Under the baseline period, no matter what the DPB a subject has during the baseline time, as long as the DBP is above 80 mHg and below 120 mmHg at the last day of baseline period, we recruit the subject into the six-month treatment phase of the trial.

In the whole simulation, we assume homogeneous variance, with a standard deviation = 8.7. The literature review showed that the standard deviation (SD) of DBP

varies from 3 to 17, with most SDs are around 7–10. So we choose 8.7 in order to make sure the SD is in the median range (Van der Lee et al. 1999; Ukena et al. 2011; Tu et al. 2000; Ford and Norrie 2002; Appel et al. 1997, 2010; Mizuno and Monteiro 2013; Nussbaumerov and Rosolov 2013; Bakx et al. 1999; Weber et al. 2013; ACCORD Study Group 2010; Hansson et al. 1998; Rdzanek et al. 2006; Kramoh et al. 2011).

2.5.2 Baseline Run-in, Pre-randomization Period

2.5.2.1 Linear Cases

We simulate baseline run-in data (prior to randomization and treatment intervention onset) that would be representative of a hypertensive cohort of patients before beginning treatment with an anti-hypertensive medication. The size of the cohort is 100 patients. We assume that their diastolic blood pressure is measured once (in practice this may be the average of 3 consecutive measurements 5 min apart) each day for 8 consecutive days at around the same time each day (say around 10:00 AM). We further assume that the DBP at the time of measurement ranges from 80 to 120 mmHg, the mean of each baseline day follows the three linear cases scenario and $SD = 8.7$. For simulation purposes, we assume the DBP distribution is normal.

Specifically, let X_{ij} denote the random variable baseline run-in DBP on the ith day of the jth patient. The baseline run-in data reflecting linearity can be simulated by assuming that means over patients follow a straight line over the baseline run-in days. That is, DBP, X_{ij}, of the jth patient on the ith day, is of the form:

$$X_{ij} = \alpha_i + \beta_i D_i + {}_{ij} \tag{2.1}$$

where $i = 1, 2, \ldots, 8$; $j = 1, 2, \ldots 100$; α_i is the intercept at day D_i, β_i is slope parameter at day D_i, and where ${}_{ij}$ follow a normal distribution with mean 0 and standard deviation σ_i. The intercepts and slope are chosen to reflect the three cases of linearity such that the X_{ij} s are constrained to the (80; 120] interval.

Case 1: The Data Are Constant Over the Number of Days

Here we assume the mean to be 95 with the standard deviation 8.7 on days = 1, 2, …, 8; $X_{ij} \sim N\ (95, \sigma_i^2)$ and $SD = 8.7$.

Case 2: The Data Reflect an Upward Trend Over the Number of Days

Here we assume the mean $E(X)$ is determined from $E(X) = 80 + \text{slope} * (Day - 1)$ for Day = 1, 2, …, 8; where $slope = (95 - 80)/7$, $X_{ij} \sim N\ (E(X), \sigma_i^2)$ and $SD = 8.7$.

Based on the different means across days, we have 8 normal distributions for observations on those days such that, for each day, the mean is increasing as compared to previous days, and starts at 80 mmHg on day 1 and steadily rises to 95 mmHg on day 8.

Case 3: The Data Reflect a Downward Trend Over the Number of Days

Here we assume the mean $E(X)$ is determined from $E(X) = 120 - slope * (Day - 1)$ for day $= 1, 2, \ldots, 8$; where $slope = (120 - 95)/7$, $X_{ij} \sim N\ (E(X), \sigma_i^2)$ and $SD = 8.7$.

Based on the different means across days, we have 8 normal distributions for observations on those days such that, for each day, the mean is decreasing as compared to previous days, and starts at 120 mmHg on day 1 and steadily drops to 95 mmHg on day 8.

For each of the three linear cases above we simulate data on 100 subjects from multivariate normal distribution (for each subject). This gives 100 simulated observations for each of 8 days, for a total of 800 simulated baseline run-in observations from each case and each simulation run. In performing the simulation, individual data points were rounded to the nearest integer and constrained to the interval (80; 120].

It should be noted constraining DBP to the (80,120] interval means that DBP chosen randomly from the distributions that lie outside the interval are rejected. This is similar to applying inclusion/exclusion criteria in a hypertension clinical trial in identifying the population to be studied. For the three linear cases, we assume the DBP of all subjects will be above or around 95 mmhg, and all subjects will be eligible to be randomized to either treatment or placebo at the end of the baseline run-in period.

2.5.2.2 Non-linear Cases

For non-linear case, we have similar assumptions. We simulate baseline run-in data for a cohort of 100 patients. We assume that their diastolic blood pressure (DBP) is measured once (in practice this may be the average of 3 consecutive measurements 5 min apart) each day for 8 consecutive days at around the same time each day (say around 10:00 AM). DBP at the time of measurement ranges from 80 to 120 mmHg. Since we have three different scenarios on each non-linear case, the baseline X_{ij} will have three different forms, respectively. We use the same standard deviation ($SD = 8.7$) as we did in the linear cases in order to maintain the same variability among data. All the baseline DBP measurements will be in the (80; 120] interval (any simulated DBP outside the interval is excluded). We still assume that the DBP distribution is normal.

Case 4: The Data Are Periodic as Reflected by the Sine Function

Here we assume the mean of DBP over the baseline run-in period oscillates as a sine function with standard deviation 8.7 on days = 1, 2, …, 8; where *slope* = sin(*day* − 1), the mean $E(X) = (a+b)/2 + ((a-b)/2) * slope$, $i = 1, 2, \ldots, 8$, $j = 1, 2, \ldots, 100$, and where $a = 80$ is the lower bound, $b = 120$ is the upper bound.

Case 5: The Data Reflect Mixed Linear and Non-linear Data

Here we use the 4 data sets that have been generated above and combine to form a new data set: randomly choose subjects from case 1 through case 4 datasets. The new dataset will have 100 subjects with 25% randomly selected from each dataset, respectively. This dataset has no specific data trend across days.

Case 6: The Data Derive from a Random Walk

Here we assume the mean of baseline period behaves as a random walk process with standard deviation of 8.7 on days = 1, 2, …, 8. The mean for each day may be any number from 80 to 120 mmHg. $X_{ij} \sim N\ (E(X), \sigma_i^2)$ and $SD = 8.7$.

2.5.3 Post Randomization Treatment Period

For the linear and the non-linear cases, we randomize all subjects at the end of the baseline run-in period into a placebo group or a treatment group in balanced fashion (50 per arm). Data are then simulated over a six-month period so that there will be 100 post-baseline DBP measurements at each month-for a total of 600 post-baseline DBPs over the six-month treatment period.

For the Placebo group, we assume mean DBP to be 90 mmHg at the end of the treatment period for all six linear and non-linear baseline cases: an average decrease of 5 mmHg in DBP over the six-month treatment period. For the Treatment group, we assume mean DBP to be 80 mmHg at the end of the treatment period for all six cases: an average decrease of 15 mmHg in DBP over the six-month treatment period. Note that the mean of the distribution of DBP just prior to randomization (after measurement on baseline run-in day 8), is around or above 95 mmHg, for six baseline run-in period cases.

Let DBP Y_{qij} denote the post-baseline run-in blood pressure measurements, q index intervention group: $q = 1$ (Placebo), 2 (Treatment), i = 1, 2, …, 6; j = 1, 2, …. The data sets for the Treatment group and the Placebo group are simulated assuming:

- $Y_{1ij} \sim N\ (\mu_{1i},\ \sigma_1^2)$, The Placebo group, and i = 1, 2, …, 6, j = 1, 2, …, n_1
- $Y_{2ij} \sim N\ (\mu_{2i},\ \sigma_1^2)$, The Treatment group, and i = 1, 2, 6, j = 1, 2, …, n_2

Where $\sigma_1 = \sigma_2 = 8.7$ and μ_{1i}=Day 8 − (5/6) ∗ i and μ_{2i}= Day 8 − (15/6) ∗ i, $i = 1, 2, 3, 4, 5, 6$.

2.5.4 Data Simulation Process

2.5.4.1 Variance-Covariance Structure

In the analysis of data from clinical trials with repeated measures, two main variance covariance structures are usually considered: autoregressive (AR(1)) and compound symmetry (CS). Since there are 14 (8 over the baseline run-in period and 6 post over the treatment period) repeated DBP measures in the 6 cases, we also consider use AR(1) and CS, with different correlation among these DBP variables. The two variance-covariance structures are:

AR(1):

$$\sum = \sigma^2 \begin{pmatrix} 1 & \cdots & \rho^{14} \\ \vdots & \ddots & \vdots \\ \rho^{14} & \cdots & 1 \end{pmatrix}$$

CS:

$$\sum = \sigma^2 \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}$$

In our cases, we have 14 repeated measures, assume the variance-covariance structures are AR(1) and CS, then, the mathematical expression is

AR(1):

$\rho = 0.9$, or $\rho = 0.5$, or $\rho = 0.1$,

$$\sum = 8.7^2 \begin{pmatrix} 1 & \cdots & \rho^{14} \\ \vdots & \ddots & \vdots \\ \rho^{14} & \cdots & 1 \end{pmatrix}$$

CS:

$\rho = 0.9$, or$\rho = 0.5$, or $\rho = 0.1$,

$$\sum = 8.7^2 \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix}$$

For the 6 cases, let $\rho = 0.9$, 0.5 and 0.1, separately, there are three AR (1)s and three CSs with the different ρ. At the end, we have 6 variance-covariance matrix for generating data under multivariate normal distribution for the 6 cases, which will be 36 datasets.

2.5.4.2 Baseline and Post-baseline Data Simulation

The DBPs are distributed as truncated multivariate normal distribution. DBP data are simulated for 100 subjects at each of the baseline period of 8 days and six-month treatment time, one DBP measurement per each subject for each day and each month.

With two kinds of variance-covariance structures (matrices), AR (1) and CS, we use the truncated multivariate normal R code package to generate the 6 cases DBP datasets, where DBPs are randomly chosen from the distributions. DBPs that lie outside the interval (80; 120] are rejected and the distributions resampled until the requisite numbers of DBPs lying in the interval are obtained. This mimics the screening of patients so that those satisfying the inclusion DBP criteria only are allowed to enter the protocol.

1. Baseline settings

The length of the baseline run-in period is 8 days. The mean and standard deviation (SD) at each day as specified below:

(1) Constant mean at each day, mean = 95, SD = 8.7, n = 100, day = 1, 2, …, 8;
(2) Increasing mean data set: $mean = 80 + k * (day - 1)$, $k = (95 - 80)/7$, $SD = 8.7$, $n = 100$, day = 1, 2, …, 8;
(3) Decreasing mean data set: $mean = 120 - k * (day - 1)$, $k = (120 - 95)/7$, $SD = 8.7$, $n = 100$, day = 1, 2, …, 8;
(4) Sine trend mean data set: $mean = (a+b)/2 + ((a - b)/2) * k$, $k = \sin(day - 1)$, $a = 80$, $b = 120$, $SD = 8.7$, $n = 100$, day = 1, 2, …, 8;
(5) Mix mean data set: choose 100 subjects from the four generated datasets; the probability of chosen from each dataset is 25%.
(6) Random walk mean data set: mean = uniform (80,120). Choose 8 means and $SD = 8.7$, $n = 100$, day = 1, 2,…, 8;

2. Post-baseline settings

The length of the treatment period is 6 months for the 6 cases-3 linear cases and 3 non-linear cases. We have six DBP measurements for each subject, one per month that we need to simulate. The 100 subjects in the baseline run-in period are randomly assigned in equal numbers to the Placebo group or to the Treatment group

just after the last observation on Day 8 (considered as Baseline for the treatment period). We simulate data over the six-month treatment period so that the average difference between the treatment group and placebo group is 10 mmHg at the end of the 6 months treatment period.

The DBP for each subject at the beginning of the treatment period is the subject's DBP at Day 8 of the baseline run-in period.

(1) Simulate Group One-Placebo

We assume that a linear average decrease of 5 mmHg in the placebo group occurs over the six-months treatment period. This is equivalent to assuming that the decrease is described by a constant slope of 5/6. That is, the mean of placebo group observations at month M is $[Day8 - (5/6)M]$ mmHg. The standard deviation at each month is assumed to be 8.7.

(2) Simulate Group Two-Treatment

We assume that a linear average decrease of 15 mmHg in the treatment group occurs over the six-months treatment period. This is equivalent to assuming that the decrease is described by a constant slope of 15/6. That is, the mean of treatment group observations at month M is $[Day8 - (15/6)M]$mmHg. The standard deviation at each month is assumed to be 8.7.

2.5.4.3 Merge Data from Baseline Run-in and Treatment Periods

We create 36 separate merged data sets that derive from the behavior of the DBP means over the baseline run-in period: constant, increasing, decreasing, sine-trend, mixed-means and random walk and the 6 variance-covariance structures. Each data set will have the variables: Subject ID No., DBP1, DBP2, DBP3, DBP4, DBP5, DBP6, DBP7, DBP8, Group, DBP9, DBP10, DBP11, DBP12, DBP13, and DBP14, where Subject ID No. is a unique subject identifier, DBP1 through DBP8 are the DBP observations on days 1 through 8 of the baseline run-in period, respectively; Group identifies the treatment period intervention group (Placebo or Treatment) into which subjects were randomized just after the day 8 observations; and DBP9 through DBP14 are the DBP observations at months 1 through 6 over the treatment period, respectively.

2.5.4.4 Compute Covariates and Final Datasets

For each of the baseline run-in data sets we identify 11 covariates as functions of the DBP observations for each subject over baseline run-in days 1 through 8. These are:

- Mean of the DBP observations over days 1 through 8 (MEAN)
- Last DBP observation-which occurs at day 8 (LO)
- First DBP observation-which occurs at day 1 (FO)

- Median of DBP observations over days 1 through 8 (MO)
- Maximum DBP observation over days 1 through 8 (MAX)
- Minimum DBP observation over days 1 through 8 (MIN)
- Area under the DBP observations time curve over days 1 through 8 (AUC)
- Average of the minimum and maximum over days 1 through 8 (Mm)
- Relative Fluctuation in DBP [(Max-Min)/Mean] over days 1 through 8 (Rate)
- Standard deviation of DBP observations over days 1 through 8 (SD)
- Coefficient of variation of DBP [Mean/SD] over days 1 through 8 (CV)

It is noted that other functions of the DBP observations over the baseline run-in period are possible. For example, for the observations generated, we could compute the mean of the observations on days 8 and 7; on days 8, 7 and 6; on days 8, 7, 6, and 5; etc. This could possibly shed some insight on how long should be the baseline run-in period.

Finally, after we compute the 11 covariates on each subject for each dataset, separately, the data sets for statistical assessment of the best covariate are obtained by adding the individual subject covariates to the data sets generated previously. The final data sets for statistical assessment thus contain the original simulated DBP data and the computed covariates and may be studied further.

2.6 Statistical Assessment of Covariates

2.6.1 Models Considered

The primary objective of this research is to identify what function of the baseline run-in DBP data is best for use as a covariate in the comparative analysis of DBP over the post-baseline run-in treatment period. To do this we consider two situations:

(1) The usual analysis of variance (ANOVA) model that includes fixed effects that capture the design sources of variation other than a covariate:

$$Y_{ijk} = \mu + \tau_i + \pi_j + (\tau * \pi)_{ij} + _{ijk} \tag{2.2}$$

where $_{ijk} \sim N(0, \sigma^2)$, Y_{ijk} is DBP of the kth subject in intervention group i at month j over the post-baseline run-in period, μ is an effect common to all subjects, τ_i is the fixed effect of intervention (Placebo or Treatment), π_j is the fixed effect of month (1, 2, 3, 4, 5, 6),$(\tau * \pi)_{ij}$ is the fixed interactive effect of intervention and time, and $_{ijk}$ is the measurement error term in observing Y_{ijk}.

(2) The analysis of covariance (ANCOVA) model that includes the fixed effects in the ANOVA model (2) but also includes a covariate:

$$Y_{ijk} = \mu + \gamma_{ik} + \tau_i + \pi_j + (\tau * \pi)_{ij} + _{ijk} \tag{2.3}$$

where $_{ijk} \sim N\ (0,\ \sigma^2)$, and in addition to the fixed effects identified in the ANOVA model (2), γ_{ik} is one of the eleven covariates (identified above) on the kth subject in intervention (Placebo or Treatment) group i (1 for Placebo, 2 for Treatment).

If a covariate has analysis utility, then the MSE from the ANCOVA model (3) has to be less than the MSE from the ANOVA model (2). Thus the best covariate to use in the comparative analysis of the post-baseline run-in data is the one that has the greatest reduction in MSE. The adjusted correlation (R ADJ) from the ANCOVA model (3) is also informative as it is a measure of the strength of the linear relationship the covariate has with the response (DBP) variable.

We considered two classifications of models used for statistical analysis of the post- baseline run-in DBP data; those with treatment-by-time interaction and those without treatment-by-time interaction. For those with treatment-by-time interaction, we consider the repeated measures framework linear model and the longitudinal data analysis model, each with and without a covariate. For those without treatment-by-time interaction, we consider the repeated measures framework linear model and the longitudinal data analysis model, each with and without a covariate. In addition, we present analysis results of DBP averaged across months for each patient (Models 23). In describing the models below, we use English rather than Greek symbolism and suppress the error term.

Class 1: Statistical Analysis Models with Treatment-by-Time interaction

- Models 11: Repeated measures analysis models of DBP
  - Model 111 : DBP = trt + time + trt * time
  - Model 112 : DBP =covariate +trt +time +trt * time
- Models 12: Longitudinal data analysis models of DBP
  - Model 121 : DBP = trt + time + trt * time
  - Model 122 : DBP = covariate + trt + time + trt * time

Class 2: Statistical Analysis Models without Treatment-by-Time interaction

- Models 21: Repeated measures analysis models of DBP
  - Model 211: DBP = trt + time
  - Model 212: DBP = covariate + trt + time

- Models 22: Longitudinal data analysis models of DBP
  - Model 221: DBP = trt + time
  - Model 222: DBP = covariate + trt + time
- Models 23: Analysis models of mean of post-DBP as response variable
  - Model 231: Mean of postDBP = trt
  - Model 232: Mean of postDBP = covariate + trt

2.6.2 Analysis Strategy

Our analysis strategy is based upon statistically analyzing simulated data sets for each of the 36 settings produced by the 6 cases of DBP mean behavior over the baseline run-in period and the 6 variance-covariance structures, using the 10 models described in the previous section, and then replicating 1000 times.

For each of these simulated data sets, we compute the 11 covariates that are functions of the baseline run-in data. We then statistically analyze each data set by running the 10 models described in the previous section and capture in tables features of the output from these analyses.

The features or analysis statistic captured in tables are: MSE from the repeated measures analysis linear model (ANOVA) without covariate, MSE for repeated measures analysis linear model (ANCOVA) with covariate (MSE1), the adjusted correlation (R ADJ) from the ANCOVA model, the AIC from the longitudinal data analysis model without covariate, and the AIC1 from the longitudinal data analysis model with covariate. Finally, we list in summary tables the means of the MSEs, the R ADJs and the AICs for the 1000 simulations.

We also capture the P-value for the treatment-by-time interaction (P_INT) from the models as a type of check on model validity. Since we simulated the data sets such that there were constant declines over the six-month treatment period in each group, with a significantly greater decline of 10 mmHg in the treatment group than in the Placebo group at 6 months, we should expect to see a statistically significant treatment-by-time interaction; i.e. at the 5% Type I error level, 95% of the P INT $<$ 0.05.

We summarize our assessment of the best choice of a covariate for the models with treatment-by-time interaction separate from the models without treatment-by-time interaction and contrast the findings from each. Readers may determine on their own whether a model with treatment-by-time interaction or a model without treatment-by-time interaction is more appropriate for their research goal as well as experimental design.

Table 2.1 Linear case covariate summary: AR(1)

Model	Best Cov.	MSE	MSE1	R.adj	AIC	AIC1
Case 1 Lin.Con						
11/12	LO	42.60	28.21	0.44	3565	3503
21/22	LO	43.49	29.22	0.42	3606	3601
23	LO	30.83	16.57	0.48	NA	NA
Case 2 Lin.Inc						
11/12	LO	43.94	29.25	0.46	3572	3512
21/22	LO	44.91	30.34	0.44	3616	3614
23	LO	32.09	17.53	0.48	NA	NA
Case 3 Lin.Dec						
11/12	LO	36.61	26.15	0.37	3520	3469
21/22	LO	37.38	27.01	0.34	3558	3557
23	LO	25.42	15.08	0.42	NA	NA

2.6.3 Analysis Results

In this section, we present the simulation results from the 6 cases under different models with separate variance-covariance structures.

This section has two subsections, linear cases and non-linear cases. In each subsection, we first present the results under AR (1) with correlation is 0.9. Secondly, we present the results under CS with correlation of 0.9. Here we only show the best covariate, which gives the smallest MSE under each case for each model. Other results and details may be found in the Appendix.

In reviewing the results presented in tabular form, the following notation and conventions are helpful:

MSE From the repeated measures analysis linear model (ANOVA) without covariate
MSE1 From the repeated measures analysis linear model (ANCOVA) with covariate
R ADJ The adjusted correlation from the ANCOVA model
AIC From the longitudinal data analysis model without covariate
AIC1 From the longitudinal data analysis model with covariate
NA Not Applicable

2.6.3.1 Linear Cases

Autoregressive AR (1) Covariance Structure

The results for AR (1) with Correlation Coefficient = 0.9 appear in the Table 2.1.

Table 2.2 Linear case covariate summary: CS

Model	Best Cov.	MSE	MSE1	R.adj	AIC	AIC1
Case 1 Lin.Con						
11/12	AUC	37.88	8.40	0.82	3245	2979
21/22	AUC	39.51	10.28	0.78	3372	3126
23	AUC	31.76	2.30	0.93	NA	NA
Case 2 Lin.Inc						
11/12	AUC	30.17	8.49	0.81	3223	2987
21/22	AUC	31.91	10.43	0.77	3353	3351
23	AUC	23.96	2.30	0.92	NA	NA
Case 3 Lin.Dec						
11/12	AUC	11.74	7.59	0.61	3035	2921
21/22	AUC	13.03	8.91	0.55	3134	3063
23	AUC	6.16	2.01	0.70	NA	NA

For the three, linear cases under AR (1), overall, the MSEs for models with interaction are smaller than the MSEs for models without interaction. With the high correlation of 0.9, the last observation (LO) is the covariate that gives the greatest reduction in MSE.

For the post mean DBP model 23, when the correlation is close to 0 (=0.1), the MSEs from all covariate models are similar and smaller than the MSEs from all covariate models for the other two correlations. This means that the mean of post-DBP has no relation with baseline. We therefore summarize in the text of this paper the results only for the highest correlation (=0.9). The other results appear in the Appendix.

Under these two types of covariance structures, the longitudinal model (class 1), with treatment-by-time interaction, leads to conclusions about the best covariate based on AICs that are similar to those from the linear models with interaction based on MSEs. For example, if the LO is the best covariate based on the linear model with smallest MSE, then LO is also the best covariate based on the smallest AIC from the longitudinal model. However, for the models without treatment-by-time interaction term, even though the ANCOVA analysis shows LO is the best covariate, MEAN is the best covariate from the longitudinal model, based upon the smallest AIC (see Appendix).

Compound Symmetry CS Covariance Structure

The results for CS with Correlation Coefficient = 0.9 appear in the Table 2.2.

For the three linear cases under CS, overall, the MSEs from the linear models with interaction are smaller than the MSEs from the linear models without interaction. When we have high correlation among the DBP over the entire trial, AUC is the covariate that gives the smallest MSE and largest adjusted R. The best covariate

Table 2.3 Linear case summary

Model	Case 1	Case 2	Case 3
Covariance: AR(1)			
11/12	LO	LO	LO
21/22	LO	LO	LO
23	No specific	No specific	No specific
Covariance: CS			
11/12	AUC(Mean)	AUC(Mean)	AUC(Mean)
21/22	AUC(Mean)	AUC(Mean)	AUC(Mean)
23	AUC(Mean)	AUC(Mean)	AUC(Mean)

determined from the models with or without the treatment-by-time interaction term are consistent with each other.

For the model with response being the mean of post-DBP, under the CS structure, AUC is the best covariate regardless of the correlation coefficient considered. Details are in the Appendix.

For the longitudinal model with treatment-by-time interaction term, the best covariates based on AICs for the three linear cases are very consistent with those from the ANCOVA models based on MSEs. The same consistency is not seen when the treatment-by-time interaction is excluded. In case 1, Max as a covariate gives the smallest AIC, so does Max in case 3 and Min in case 2. The Appendix may be seen for details.

Linear Cases Summary

We have assessed the best covariate for the three linear cases of baseline run-in data using the analysis models of treatment period data assuming two covariance structures with three levels of correlation ($\rho = 0.1$, 0.5 and 0.9). For smaller values of correlation there is no clear advantage of one covariate over another (see Appendix). The Table 2.3 summarizes the assessment of the best covariate when the DBP data are highly correlated ($\rho = 0.9$).

Summary of Best function of baseline as covariates for 3 linear cases in Table 2.3. From this table the findings are:

1. Under Models 11/12 and 21/22 and AR (1) with high correlation, the last observation (LO) is the best covariate. For Model 22, the longitudinal models without interaction term, mean gives the smallest AIC, which is not consistent with the corresponding ANCOVA model.
2. Under Models 11/12 and 21/22 and CS with high correlation, AUC (or MEAN as its value is close to that of AUC because the single daily observations taken over the baseline run-in period were a day apart) is the best covariate and has the larger adjusted R Square.

Table 2.4 Non-Linear case covariate summary: AR(1)

Model	Best Cov.	MSE	MSE1	R.adj	AIC	AIC1
Case 4 Non-Lin.Sine						
11/12	LO	24.13	22.61	0.17	3387	3378
21/22	LO	24.59	23.09	0.15	3412	3413
23	LO	14.58	13.16	0.17	NA	NA
Case 5 Non-Lin.Mix						
11/12	MAX	36.51	32.10	0.23	3704	3648
21/22	MAX	37.19	32.82	0.21	3720	3696
23	MAX	25.36	21.12	0.20	NA	NA
Case 6 Non-Lin.RW						
11/12	LO	33.05	25.79	0.81	3511	3472
21/22	LO	86.27	79.16	0.45	4191	4190
23	LO	21.65	14.50	0.49	NA	NA

For Model 22, the longitudinal models without interaction term, MAX as a covariate gives the smallest AIC in case 1, MAX in case 3 also gives the smallest AIC, while MIN in case 2 gives the smallest AIC, they are not consistent with the corresponding ANCOVA models. (see Appendix).

3. Under Model 23 for post-mean DBP with AR (1), no covariate is preferred to another.
4. Under Model 23 with CS, AUC is the best covariate.

2.6.3.2 Non-linear Cases

Autoregressive AR (1) Covariance Structure

The results for AR (1) with Correlation Coefficient = 0.9 appear in the Table 2.4.

For the three, non-linear cases under AR (1), overall, the MSEs from the models with treatment-by-time interaction are smaller than the MSEs from the models without treatment- by-time interaction; the greater is the correlation, the smaller is the MSE and the larger is the Adjusted R square. The last observation (LO) as covariate gives the smallest MSE and the largest adjusted R Square in non-linear cases 4 and 6, and the maximum (MAX) as a covariate gives the smallest MSE and largest adjusted R Square in non-linear case 5.

For model 23 where the mean of post baseline DBP is the response variable, we observe the same results for the non-linear cases as we did for the linear cases under AR (1). Correlation and MSE behave monotonically; i.e. the smallest correlation leads to the smallest MSE and the largest correlation leads to the largest MSE. Further no covariate is preferred to another and the reduction in MSE by adding any covariate to the ANOVA model is near zero.

Table 2.5 Non-linear case covariate summary: AR(1) special case

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	86.27	0.40	86.09	0.35	76.43	0.34	76.43
Mean	82.98	0.42	86.00	0.35	76.46	0.34	76.46
LO	79.16	0.45	85.56	0.35	76.47	0.34	76.47

For non-linear cases and AR (1), the results from the longitudinal model with treatment- by-time interaction are consistent with those from their corresponding ANCOVA models, respectively. The best covariate from ANCOVA has the smallest AIC. For the longitudinal model without interaction term, the results are not consistent with the results from ANCOVA model.

Actually, under AR(1), in model 11/12, model 21/22 and model 23, no matter what the correlation is, the MSEs within the same correlation are similar-which says that no covariate improves precision. The MSEs from LO and MAX are relatively smaller among all covariates according to each case.

Specifically, for the random walk case 6 and the model without interaction term, correlation and MSEs behave monotonically. Regardless of the size of the correlation, MSEs within the same correlation are similar. Further, for the smallest correlation, adding a covariate to the ANOVA model virtually has no effect on the MSE. This result seems opposite to results from model 11/12 and model 22 see Table 2.5.

Non-linear Case 6 for model without treatment-by-time interaction. In Table 2.5.

Compound Symmetry (CS) Covariance Structure

The results from assuming compound symmetry covariance structure with correlation coefficient = 0.9 appear in the Table 2.6.

For the three, non-linear cases under CS, overall, the MSEs from the models with inter- action are smaller than the MSEs from the models without interaction, and the greater is the correlation, the smaller is the MSE and the larger is the adjusted R square. In model 11/12, model 21/22 and model 23, no matter what the correlation is, the MSEs within the same correlation are similar which says that no covariate improves precision. The AUC or Mean as covariates give the smallest MSE and largest adjusted R Square for non-linear cases 4 and 6, and the maximum (MAX) as a covariate gives the smallest MSE and largest adjusted R Square for non-linear case 5.

Table 2.6 Non-linear case covariate summary: CS

Model	Best Cov.	MSE	MSE1	R.adj	AIC	AIC1
Case 4 Non-Lin.Sine						
11/12	AUC	9.42	7.80	0.63	2996	2938
21/22	AUC	10.71	9.10	0.57	3092	3068
23	AUC	3.68	2.06	0.73	NA	NA
Case 5 Non-Lin.Mix						
11/12	MAX	26.80	17.70	0.51	3477	3306
21/22	MAX	28.19	19.18	0.47	3531	3467
23	MAX	20.19	11.90	0.45	NA	NA
Case 6 Non-Lin.RW						
11/12	AUC	24.86	8.27	0.94	3186	2972
21/22	AUC	79.76	63.40	0.56	3372	2979
23	AUC	18.79	2,21	0.91	NA	NA

Table 2.7 Non-linear case summary

Model	Case 4	Case 5	Case 6
Covariance: AR(1)			
11/12	LO	MAX	LO
21/22	LO	MAX	LO
23	LO	MAX	LO
Covariance: CS			
11/12	AUC(Mean)	MAX	AUC(Mean)
21/22	AUC(Mean)	MAX	AUC(Mean)
23	AUC(Mean)	MAX	AUC(Mean)

For non-linear longitudinal model with interaction term, the results are consistent with their corresponding ANCOVA models, respectively. The best covariate producing the smallest MSE also produces the smallest AIC. For the longitudinal model without interaction term, best covariate results are inconsistent between the ANCOVA model and the longitudinal model (see Appendix).

Non-linear Cases Summary

We have assessed the best covariate for the three non-linear cases of baseline run-in data using the analysis models of treatment period data assuming two covariance structures with three levels of correlation ($\rho = 0.1$, 0.5 and 0.9). For smaller values of correlation there is no clear advantage of one covariate over another (see Appendix). The Table 2.7 summarizes the assessment of the best covariate when the DBP data are highly correlated ($\rho = 0.9$).

Summary of Best function of baselines as covariates for non-linear cases. In Table 2.7.

From this table the findings for the non-linear cases are:

1. The results are dissimilar across the three non-linear cases. Cases 4 and 6 yield the same results while case 5 yields different results.
2. For case 4 and 6, under AR (1) the best covariate is LO; and under CS, the best covariate is AUC (or Mean).
3. For case 5, with high correlation under AR (1) and CS, MAX is the best covariate function of baseline and gives the smallest MSE.
4. For post-mean DBP as response under AR(1), all MSEs are similar under each correlation, even though the MSE from LO is relatively smaller. It appears that ANCOVA produces the same results as ANOVA when analyzing the mean of Post-baseline DBP for all covariates considered. For the mean of Post-baseline DBP as response under CS, the best covariate is AUC (or Mean) in case 4 and 6. In case 5, Max is the best covariate.

2.7 Conclusions

Overall, the best function of 8 days of baseline run-in DBP for use as a covariate in the statistical analysis of post-randomization DBP treatment effect data depends on the pattern of the baseline run-in data, the strength of the correlation among DBP measurements over time and the covariance structure of these observations. For highly correlated DBP data in general, (1) the last observation (LO) in the baseline run-in period is the best covariate under AR (1) and (2) the area under the DBP-by-day curve (AUC) is the best covariate under CS.

There are practical implications from these findings regarding the design of a comparative efficacy trial in which covariance analysis of the efficacy data are planned. Having multiple days of baseline run-in for the purpose of observing the primary efficacy variable prior to randomization is not justified-if one can document that the data exhibit high correlation and an AR (1) covariance structure; rather just take one measurement prior to randomization. In the absence of such documentation, using AUC based on multiple baseline measurements is expected to lead to covariate efficiency gains beyond a single measurement prior to randomization.

Appendix

This section includes all the result tables that we did not include in the body of the chapter. A copy of the Appendix may be obtained by request from the first author.

The models used are:

Class 1: Statistical Analysis Models with Treatment-by-Time interaction

- Models 11: Repeated measures analysis models of DBP
  - Model 111 : DBP = trt + time + trt ∗ time
  - Model 112 : DBP = covariate + trt + time + trt ∗ time
- Models 12: Longitudinal data analysis models of DBP
  - Model 121: DBP = trt + time + trt*time
  - Model 122: DBP = covariate + trt + time + trt*time

Class 2: Statistical Analysis Models without Treatment-by-Time interaction

- Models 21: Repeated measures analysis models of DBP
  - Model 211: DBP = trt + time
  - Model 212: DBP = covariate + trt + time
- Models 22: Longitudinal data analysis models of DBP
  - Model 221: DBP = trt + time
  - Model 222: DBP = covariate + trt + time
- Models 23: Analysis models of mean of post-DBP as response variable
  - Model 231: Mean of postDBP = trt
  - Model 232: Mean of postDBP = covariate + trt

See Tables 2.8, 2.9, 2.10, 2.11, 2.12, 2.13, 2.14, 2.15, 2.16, 2.17, 2.18, 2.19, 2.20, 2.21, 2.22, 2.23, 2.24, 2.25, 2.26, 2.27, 2.28, 2.29, 2.30, 2.31, 2.32, 2.33, 2.34, 2.35, 2.36, 2.37, 2.38, 2.39, 2.40, 2.41, 2.42, 2.43, 2.44, 2.45, 2.46, 2.47, 2.48, 2.49, 2.50, 2.51, 2.52, 2.53, 2.54, 2.55, 2.56, 2.57, 2.58, 2.59, 2.60, 2.61, 2.62, 2.63, 2.64, 2.65, 2.66 and 2.67.

Table 2.8 Case 1 Model 21 (without interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	43.49	0.14	50.26	0.09	50.06	0.10	43.49
Mean	35.51	0.30	50.06	0.10	50.04	0.10	35.51
LO	29.22	0.42	49.40	0.11	50.04	0.10	29.22
FO	41.49	0.18	50.19	0.10	50.04	0.10	41.49
Median	35.99	0.29	50.08	0.10	50.04	0.10	35.99
MIN	35.32	0.30	50.09	0.10	50.04	0.10	35.32
MAX	36.90	0.27	50.11	0.10	50.05	0.10	36.9
AUC	35.9	0.29	50.11	0.1	50.04	0.1	35.9
Mm	35.25	0.3	50.07	0.1	50.04	0.1	35.25
Rate	43.17	0.15	50.18	0.1	50.04	0.1	43.17
SD	42.88	0.15	50.18	0.1	50.04	0.1	42.88
CV	43.17	0.15	50.18	0.1	50.04	0.1	43.17
Total—Min	29.22		49.4		50.04		29.22

Table 2.9 Case 1 Model 22 (longitudinal without interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3606	4028	4059	3606
Mean	3544	4024	4060	3544
LO	3601	4030	4060	3601
FO	3580	4029	4060	3580
Median	3577	4029	4060	3577
MIN	3584	4029	4060	3584
MAX	3579	4029	4060	3579
AUC	3576	4029	4060	3576
Mm	3607	4029	4060	3607
Rate	3606	4029	4060	3606
SD	3607	4029	4060	3607
CV	3580	4029	4060	3580
Total—Min	3544	4024	4059	3544

Table 2.10 Case 1 Model 21 (without interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	39.51	0.17	47.48	0.13	51.66	0.11	39.51
Mean	10.30	0.78	36.92	0.32	50.66	0.12	10.3
LO	15.13	0.68	43.28	0.21	51.47	0.11	15.13
FO	15.23	0.68	43.20	0.21	51.46	0.11	15.23
Median	10.54	0.78	37.60	0.31	50.82	0.12	10.54
MIN	11.87	0.75	40.08	0.27	51.18	0.11	11.87
MAX	11.87	0.75	39.98	0.27	51.22	0.11	11.87
AUC	10.28	0.78	37	0.32	50.68	0.12	10.28
Mm	10.78	0.77	38.16	0.3	50.94	0.12	10.78
Rate	38.28	0.2	47.16	0.14	51.64	0.11	38.28
SD	39.26	0.18	47.37	0.14	51.62	0.11	39.26
CV	38.11	0.2	47.13	0.14	51.64	0.11	38.11
Total—Min	10.28		36.92		50.66		10.28

Table 2.11 Case 1 Model 22 (longitudinal without interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3372	3949	4075	3372
Mean	3224	3924	4074	3224
LO	3225	3925	4074	3225
FO	3117	3877	4068	3117
Median	3161	3901	4071	3161
MIN	3159	3900	4072	3159
MAX	3105	3870	4066	3105
AUC	3126	3882	4069	3126
Mm	3369	3949	4076	3369
Rate	3373	3950	4075	3373
SD	3369	3949	4076	3369
CV	3117	3877	4068	3117
Total—Min	3105	3869	4066	3105

Table 2.12 Case 1 Model 23 AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	30.83	0.04	16.10	0.11	9.35	0.23	9.35
Mean	22.96	0.28	15.98	0.12	9.35	0.23	9.35
LO	16.57	0.48	15.30	0.15	9.34	0.23	9.34
FO	29.03	0.09	16.11	0.11	9.35	0.23	9.35
Median	23.44	0.27	16.00	0.12	9.34	0.23	9.34
MIN	22.76	0.29	16.01	0.12	9.35	0.23	9.35
MAX	24.38	0.24	16.03	0.11	9.35	0.23	9.35
AUC	23.35	0.27	16.03	0.11	9.35	0.23	9.35
Mm	22.7	0.29	15.98	0.12	9.35	0.23	9.35
Rate	30.75	0.04	16.1	0.11	9.34	0.23	9.34
SD	30.45	0.05	16.1	0.11	9.34	0.23	9.34
CV	30.75	0.04	16.1	0.11	9.34	0.23	9.34
Total—Min	16.57		15.3		9.34		9.34

Table 2.13 Case 1 Model 23 CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	31.76	0.02	19.26	0.07	10.89	0.19	10.89
Mean	2.32	0.93	8.63	0.58	9.89	0.26	2.32
LO	7.23	0.77	15.10	0.27	10.72	0.20	7.23
FO	7.34	0.77	15.02	0.28	10.71	0.20	7.34
Median	2.57	0.92	9.33	0.55	10.06	0.25	2.57
MIN	3.91	0.88	11.85	0.43	10.42	0.22	3.91
MAX	3.91	0.88	11.75	0.43	10.46	0.22	3.91
AUC	2.3	0.93	8.71	0.58	9.92	0.26	2.3
Mm	2.81	0.91	9.89	0.52	10.17	0.24	2.81
Rate	30.78	0.05	19.05	0.08	10.89	0.19	10.89
SD	31.77	0.02	19.26	0.07	10.87	0.19	10.87
CV	30.6	0.05	19.02	0.09	10.89	0.19	10.89
Total—Min	2.3		8.63		9.89		2.3

Table 2.14 Case 1 Model 11 (with interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	42.60	0.16	49.88	0.10	49.62	0.11	42.6
Mean	34.55	0.32	49.68	0.11	49.60	0.11	34.55
LO	28.21	0.44	49.00	0.12	49.60	0.11	28.21
FO	40.58	0.20	49.80	0.10	49.61	0.11	40.58
Median	35.03	0.31	49.70	0.10	49.60	0.11	35.03
MIN	34.35	0.32	49.71	0.10	49.60	0.11	34.35
MAX	35.95	0.29	49.73	0.10	49.61	0.11	35.95
AUC	34.94	0.31	49.73	0.1	49.61	0.11	34.94
Mm	34.29	0.32	49.68	0.11	49.61	0.11	34.29
Rate	42.28	0.16	49.79	0.1	49.6	0.11	42.28
SD	41.98	0.17	49.79	0.1	49.6	0.11	41.98
CV	42.28	0.16	49.8	0.1	49.6	0.11	42.28
Total—Min	28.21		49		49.6		28.21

Table 2.15 Case 1 Model 12 (longitudinal with interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3565	4025	4059	3565
Mean	3536	4025	4059	3536
LO	3503	4021	4059	3503
FO	3560	4026	4059	3560
Median	3538	4026	4059	3538
MIN	3535	4026	4059	3535
MAX	3542	4026	4059	3542
AUC	3538	4026	4059	3538
Mm	3535	4025	4059	3535
Rate	3566	4026	4059	3566
SD	3565	4026	4059	3565
CV	3566	4026	4059	3566
Total—Min	3503	4021	4059	3503

Table 2.16 Case 1 Model 11 (with interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	37.88	0.21	46.33	0.15	51.04	0.12	37.88
Mean	8.43	0.82	35.68	0.35	50.03	0.13	8.43
LO	13.30	0.72	42.09	0.23	50.85	0.12	13.3
FO	13.40	0.72	42.01	0.23	50.84	0.12	13.4
Median	8.67	0.82	36.37	0.34	50.20	0.13	8.67
MIN	10.01	0.79	38.86	0.29	50.56	0.13	10.01
MAX	10.01	0.79	38.76	0.29	50.59	0.12	10.01
AUC	8.4	0.82	35.76	0.35	50.06	0.13	8.4
Mm	8.91	0.81	36.92	0.32	50.31	0.13	8.91
Rate	36.65	0.23	46	0.16	51.02	0.12	36.65
SD	37.63	0.21	46.21	0.16	51	0.12	37.63
CV	36.47	0.24	45.98	0.16	51.02	0.12	36.47
Total—Min	8.4		35.68		50.03		8.4

Table 2.17 Case 1 Model 12 (longitudinal with interaction) CS

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3245	3933	4072	3245
Mean	2981	3853	4064	2981
LO	3097	3910	4072	3097
FO	3098	3909	4072	3098
Median	2991	3861	4065	2991
MIN	3034	3885	4069	3034
MAX	3034	3884	4069	3034
AUC	2979	3854	4064	2979
Mm	3000	3867	4066	3000
Rate	3243	3933	4073	3243
SD	3246	3934	4073	3246
CV	3242	3933	4073	3242
Total—Min	2979	3853	4064	2979

Table 2.18 Case 2 Model 21 (without interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	44.91	0.18	50.40	0.10	50.00	0.10	44.91
Mean	36.80	0.33	50.15	0.10	49.99	0.10	36.8
LO	30.34	0.44	49.52	0.11	49.98	0.10	30.34
FO	43.82	0.20	50.33	0.10	49.99	0.10	43.82
Median	38.60	0.29	50.24	0.10	50.00	0.10	38.6
MIN	32.17	0.41	50.01	0.11	49.99	0.10	32.17
MAX	42.75	0.22	50.32	0.10	49.99	0.10	42.75
AUC	37.37	0.32	50.23	0.1	49.99	0.1	37.37
Mm	35.4	0.35	50.07	0.1	49.99	0.1	35.4
Rate	39.14	0.29	50.1	0.1	49.99	0.1	39.14
SD	38.06	0.3	50.07	0.1	49.99	0.1	38.06
CV	39.44	0.28	50.1	0.1	49.99	0.1	39.44
Total-Min	30.34		49.52		49.98		30.34

Table 2.19 Case 2 Model 22 (longitudinal without interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3616	4029	4056	3616
Mean	3556	4025	4057	3556
LO	3614	4030	4057	3614
FO	3595	4030	4057	3595
Median	3566	4028	4057	3566
MIN	3590	4030	4057	3590
MAX	3582	4029	4057	3582
AUC	3597	4029	4057	3597
Mm	3593	4029	4057	3593
Rate	3598	4029	4057	3598
SD	3595	4030	4057	3595
CV	3573	4027	4056	3573
Total-Min	3556	4025	4056	3566

Table 2.20 Case 2 Model 21 (without interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	31.91	0.30	48.15	0.18	54.06	0.12	31.91
Mean	10.46	0.77	39.16	0.33	53.19	0.13	10.46
LO	15.04	0.67	44.72	0.24	53.86	0.12	15.04
FO	15.29	0.66	45.49	0.22	53.92	0.12	15.29
Median	11.55	0.75	40.26	0.31	53.36	0.13	11.55
MIN	13.47	0.70	42.72	0.27	53.67	0.12	13.47
MAX	13.56	0.70	43.49	0.26	53.79	0.12	13.56
AUC	10.43	0.77	39.26	0.33	53.21	0.13	10.43
Mm	11.69	0.74	41.06	0.3	53.53	0.13	11.69
Rate	31.44	0.31	47.84	0.18	53.95	0.12	31.44
SD	31.37	0.31	47.33	0.19	53.9	0.12	31.37
CV	31.22	0.32	47.9	0.18	53.96	0.12	31.22
Total—Min	10.43		39.16		53.19		10.43

Table 2.21 Case 2 Model 22 (longitudinal without interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3353	3976	4103	3351
Mean	3228	3956	4103	3228
LO	3231	3962	4103	3231
FO	3155	3920	4098	3155
Median	3202	3942	4101	3202
MIN	3114	3910	4097	3114
MAX	3160	3928	4099	3160
AUC	3351	3976	4103	3351
Mm	3351	3973	4103	3351
Rate	3350	3976	4104	3350
SD	3152	3920	4098	3152
CV	3222	3957	4100	3222
Total—Min	3114	3909	4096	3114

Table 2.22 Case 2 Model 23 AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	32.09	0.05	16.21	0.11	9.33	0.23	9.33
Mean	24.10	0.28	16.03	0.12	9.33	0.23	9.33
LO	17.53	0.48	15.38	0.15	9.32	0.23	9.32
FO	31.23	0.08	16.21	0.11	9.33	0.23	9.33
Median	25.93	0.23	16.12	0.11	9.33	0.23	9.33
MIN	19.39	0.42	15.88	0.13	9.32	0.23	9.32
MAX	30.15	0.11	16.20	0.11	9.32	0.23	9.32
AUC	24.68	0.27	16.11	0.11	9.33	0.23	9.33
Mm	22.68	0.33	15.95	0.12	9.32	0.23	9.32
Rate	26.47	0.21	15.97	0.12	9.32	0.23	9.32
SD	25.38	0.25	15.95	0.12	9.32	0.23	9.32
CV	26.78	0.21	15.97	0.12	9.32	0.23	9.32
Total—Min	17.53		15.38		9.32		9.32

Table 2.23 Case 2 Model 23 CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	23.96	0.20	18.21	0.18	11.36	0.21	11.36
Mean	2.33	0.92	9.16	0.58	10.50	0.27	2.33
LO	6.99	0.76	14.82	0.33	11.18	0.22	6.99
FO	7.26	0.75	15.60	0.29	11.24	0.22	7.26
Median	3.45	0.88	10.28	0.53	10.68	0.26	3.45
MIN	5.40	0.82	12.78	0.42	10.99	0.24	5.4
MAX	5.50	0.81	13.56	0.38	11.11	0.23	5.5
AUC	2.3	0.92	9.26	0.58	10.52	0.27	2.3
Mm	3.59	0.88	11.09	0.5	10.84	0.25	3.59
Rate	23.67	0.21	17.99	0.19	11.28	0.22	11.28
SD	23.6	0.21	17.47	0.21	11.22	0.22	11.22
CV	23.44	0.22	18.05	0.18	11.29	0.22	11.29
Total—Min	2.3		9.16		10.5		2.3

Table 2.24 Case 2 Model 11 (with interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	43.94	0.20	50.05	0.10	49.59	0.11	43.94
Mean	35.76	0.35	49.79	0.11	49.58	0.11	35.76
LO	29.25	0.46	49.15	0.12	49.57	0.11	29.25
FO	42.83	0.22	49.97	0.11	49.58	0.11	42.83
Median	37.57	0.31	49.88	0.11	49.59	0.11	37.57
MIN	31.09	0.43	49.65	0.11	49.58	0.11	31.09
MAX	41.76	0.24	49.96	0.11	49.58	0.11	41.76
AUC	36.34	0.34	49.87	0.11	49.59	0.11	36.34
Mm	34.35	0.37	49.71	0.11	49.58	0.11	34.35
Rate	38.11	0.3	49.74	0.11	49.58	0.11	38.11
SD	37.03	0.32	49.71	0.11	49.58	0.11	37.03
CV	38.42	0.3	49.74	0.11	49.58	0.11	38.42
Total—Min	29.25		49.15		49.57		29.25

Table 2.25 Case 2 Model 12 (longitudinal with interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3572	4027	4058	3572
Mean	3544	4027	4059	3544
LO	3512	4023	4059	3512
FO	3570	4028	4059	3570
Median	3552	4027	4059	3552
MIN	3523	4026	4059	3523
MAX	3567	4028	4059	3567
AUC	3547	4027	4059	3547
Mm	3538	4026	4059	3538
Rate	3554	4027	4059	3554
SD	3550	4026	4059	3550
CV	3555	4027	4059	3555
Total—Min	3512	4023	4058	3512

Table 2.26 Case 2 Model 11 (with interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	30.17	0.34	46.84	0.20	53.36	0.13	30.17
Mean	8.53	0.81	37.77	0.35	52.49	0.14	8.53
LO	13.14	0.71	43.38	0.26	53.16	0.13	13.14
FO	13.40	0.71	44.15	0.25	53.23	0.13	13.4
Median	9.63	0.79	38.88	0.34	52.66	0.14	9.63
MIN	11.57	0.75	41.36	0.29	52.97	0.13	11.57
MAX	11.66	0.74	42.13	0.28	53.10	0.13	11.66
AUC	8.49	0.81	37.87	0.35	52.51	0.14	8.49
Mm	9.77	0.78	39.69	0.32	52.83	0.14	9.77
Rate	29.69	0.35	46.53	0.21	53.26	0.13	29.69
SD	29.62	0.35	46.01	0.21	53.21	0.13	29.62
CV	29.46	0.36	46.58	0.2	53.27	0.13	29.46
Total—Min	8.49		37.77		52.49		8.49

Table 2.27 Case 2 Model 12 (longitudinal with interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3223	3955	4099	3223
Mean	2988	3887	4092	2988
LO	3100	3936	4098	3100
FO	3104	3941	4099	3104
Median	3028	3899	4094	3028
MIN	3074	3921	4097	3074
MAX	3076	3927	4098	3076
AUC	2987	3888	4092	2987
Mm	3032	3907	4095	3032
Rate	3223	3955	4099	3223
SD	3223	3952	4099	3223
CV	3222	3956	4099	3222
Total—Min	2987	3887	4092	2987

Table 2.28 Case 3 Model 21 (without interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5	0.1			
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	37.38	0.10	49.99	0.09	50.04	0.10	37.38
Mean	31.82	0.23	49.75	0.10	50.03	0.10	31.82
LO	27.01	0.34	49.14	0.11	50.02	0.10	27.01
FO	36.79	0.11	49.90	0.09	50.03	0.10	36.79
Median	33.46	0.19	49.84	0.09	50.03	0.10	33.46
MIN	36.45	0.12	49.90	0.09	50.03	0.10	36.45
MAX	27.94	0.32	49.56	0.10	50.02	0.10	27.94
AUC	32.24	0.22	49.82	0.09	50.03	0.1	32.24
Mm	30.45	0.26	49.64	0.1	50.03	0.1	30.45
Rate	30.93	0.25	49.63	0.1	50.02	0.1	30.93
SD	32.17	0.22	49.65	0.1	50.02	0.1	32.17
CV	31.4	0.24	49.63	0.1	50.02	0.1	31.4
Total—Min	27.01		49.14		50.02		27.01

Table 2.29 Case 3 Model 22 (longitudinal without interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3558	4024	4058	3558
Mean	3506	4020	4059	3506
LO	3557	4025	4059	3557
FO	3542	4024	4059	3542
Median	3555	4025	4059	3555
MIN	3512	4022	4059	3512
MAX	3536	4024	4059	3536
AUC	3527	4023	4059	3527
Mm	3530	4023	4059	3530
Rate	3536	4023	4059	3536
SD	3532	4023	4059	3532
CV	3542	4024	4059	3542
Total—Min	3506	4020	4058	3506

Table 2.30 Case 3 Model 21 (without interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	13.03	0.34	33.16	0.13	47.49	0.10	13.03
Mean	8.94	0.55	30.36	0.20	46.84	0.11	8.94
LO	10.88	0.45	32.30	0.15	47.34	0.10	10.88
FO	11.40	0.42	32.73	0.14	47.41	0.10	11.4
Median	9.81	0.50	31.00	0.19	47.00	0.11	9.81
MIN	10.99	0.44	32.34	0.15	47.35	0.10	10.99
MAX	10.58	0.46	31.84	0.16	47.23	0.10	10.58
AUC	8.91	0.55	30.4	0.2	46.86	0.11	8.91
Mm	9.95	0.49	31.43	0.17	47.14	0.11	9.95
Rate	12.08	0.39	32.4	0.15	47.33	0.1	12.08
SD	12.67	0.36	32.69	0.14	47.39	0.1	12.67
CV	12.15	0.38	32.43	0.15	47.34	0.1	12.15
Total—Min	8.91		30.36		46.84		8.91

Table 2.31 Case 3 Model 22 (longitudinal without interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3134	3789	4025	3134
Mean	3091	3781	4025	3091
LO	3103	3786	4026	3103
FO	3058	3765	4021	3058
Median	3093	3782	4025	3093
MIN	3083	3776	4024	3083
MAX	3019	3757	4020	3019
AUC	3063	3771	4023	3063
Mm	3118	3782	4025	3118
Rate	3129	3786	4025	3129
SD	3120	3783	4025	3120
CV	3056	3765	4021	3056
Total—Min	3019	3756	4019	3019

Table 2.32 Case 3 Model 23 AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	25.42	0.03	15.99	0.11	9.34	0.23	9.34
Mean	19.96	0.24	15.83	0.12	9.34	0.23	9.34
LO	15.08	0.42	15.21	0.15	9.33	0.23	9.33
FO	25.02	0.05	15.99	0.11	9.34	0.23	9.34
Median	21.63	0.17	15.92	0.11	9.34	0.23	9.34
MIN	24.67	0.06	15.98	0.11	9.34	0.23	9.34
MAX	16.02	0.39	15.64	0.13	9.33	0.23	9.33
AUC	20.39	0.22	15.9	0.12	9.34	0.23	9.34
Mm	18.57	0.29	15.72	0.13	9.34	0.23	9.34
Rate	19.05	0.27	15.71	0.13	9.33	0.23	9.33
SD	20.32	0.22	15.73	0.12	9.33	0.23	9.33
CV	19.53	0.25	15.71	0.13	9.33	0.23	9.33
Total—Min	15.08		15.21		9.33		9.33

Table 2.33 Case 3 Model 23 CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	6.16	0.08	9.50	0.10	9.78	0.17	6.16
Mean	2.04	0.69	6.70	0.36	9.13	0.23	2.04
LO	4.01	0.40	8.67	0.17	9.65	0.18	4.01
FO	4.54	0.32	9.11	0.13	9.72	0.18	4.54
Median	2.92	0.56	7.35	0.30	9.30	0.21	2.92
MIN	4.12	0.38	8.70	0.17	9.65	0.18	4.12
MAX	3.71	0.44	8.21	0.22	9.53	0.19	3.71
AUC	2.01	0.7	6.73	0.36	9.15	0.23	2.01
Mm	3.07	0.54	7.79	0.26	9.43	0.2	3.07
Rate	5.23	0.22	8.77	0.17	9.63	0.19	5.23
SD	5.83	0.13	9.06	0.14	9.69	0.18	5.83
CV	5.3	0.21	8.8	0.16	9.64	0.19	5.3
Total—Min	2.01		6.7		9.13		2.01

Table 2.34 Case 3 Model 11 (with interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5	0.1			
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	36.61	0.12	49.62	0.10	49.62	0.11	36.61
Mean	30.99	0.25	49.38	0.10	49.61	0.11	30.99
LO	26.15	0.37	48.77	0.11	49.60	0.11	26.15
FO	36.01	0.13	49.54	0.10	49.61	0.11	36.01
Median	32.66	0.21	49.47	0.10	49.61	0.11	32.66
MIN	35.67	0.14	49.53	0.10	49.61	0.11	35.67
MAX	27.09	0.34	49.19	0.11	49.60	0.11	27.09
AUC	31.43	0.24	49.45	0.1	49.61	0.11	31.43
Mm	29.62	0.28	49.27	0.1	49.61	0.11	29.62
Rate	30.1	0.27	49.26	0.1	49.6	0.11	30.1
SD	31.35	0.24	49.28	0.1	49.6	0.11	31.35
CV	30.58	0.26	49.26	0.1	49.6	0.11	30.58
Total—Min	26.15		48.77		49.6		26.15

Table 2.35 Case 3 Model 12 (longitudinal with interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3520	4022	4059	3520
Mean	3497	4022	4060	3497
LO	3469	4018	4059	3469
FO	3520	4023	4059	3520
Median	3505	4023	4060	3505
MIN	3518	4023	4059	3518
MAX	3475	4021	4059	3475
AUC	3499	4023	4060	3499
Mm	3490	4022	4059	3490
Rate	3492	4022	4059	3492
SD	3499	4022	4059	3499
CV	3495	4022	4059	3495
Total—Min	3469	4018	4059	3469

Table 2.36 Case 3 Model 11 (with interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	11.74	0.40	32.42	0.15	46.99	0.11	11.74
Mean	7.62	0.61	29.60	0.22	46.33	0.12	7.62
LO	9.57	0.51	31.56	0.17	46.84	0.11	9.57
FO	10.10	0.49	31.99	0.16	46.91	0.11	10.1
Median	8.49	0.57	30.25	0.21	46.50	0.12	8.49
MIN	9.68	0.51	31.59	0.17	46.84	0.11	9.68
MAX	9.27	0.53	31.10	0.18	46.73	0.11	9.27
AUC	7.59	0.61	29.64	0.22	46.35	0.12	7.59
Mm	8.64	0.56	30.68	0.19	46.63	0.12	8.64
Rate	10.78	0.45	31.66	0.17	46.83	0.11	10.78
SD	11.38	0.42	31.95	0.16	46.89	0.11	11.38
CV	10.85	0.45	31.69	0.17	46.84	0.11	10.85
Total—Min	7.59		29.6		46.33		7.59

Table 2.37 Case 3 Model 12 (longitudinal with interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3035	3779	4023	3035
Mean	2923	3744	4018	2923
LO	2992	3771	4023	2992
FO	3005	3776	4024	3005
Median	2960	3754	4019	2960
MIN	2995	3771	4023	2995
MAX	2984	3765	4022	2984
AUC	2921	3745	4018	2921
Mm	2965	3760	4021	2965
Rate	3019	3772	4023	3019
SD	3030	3775	4024	3030
CV	3021	3772	4023	3021
Total—Min	2921	3744	4018	2921

Table 2.38 Case 4 Model 21 (without interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5	0.1			
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	24.59	0.09	46.96	0.06	49.65	0.09	24.59
Mean	24.26	0.11	46.87	0.06	49.64	0.09	24.26
LO	23.09	0.15	46.69	0.07	49.64	0.09	23.09
FO	24.47	0.10	46.90	0.06	49.64	0.09	24.47
Median	24.39	0.10	46.88	0.06	49.64	0.09	24.39
MIN	23.95	0.12	46.85	0.06	49.64	0.09	23.95
MAX	24.46	0.10	46.89	0.06	49.64	0.09	24.46
AUC	24.31	0.1	46.88	0.06	49.64	0.09	24.31
Mm	24.2	0.11	46.88	0.06	49.64	0.09	24.2
Rate	24.44	0.1	46.89	0.06	49.64	0.09	24.44
SD	24.38	0.1	46.88	0.06	49.64	0.09	24.38
CV	24.44	0.1	46.89	0.06	49.64	0.09	24.44
Total—Min	23.09		46.69		49.64		23.09

Table 2.39 Case 4 Model 22 (longitudinal without interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3412	3987	4055	3412
Mean	3403	3987	4056	3403
LO	3413	3988	4056	3413
FO	3412	3988	4056	3412
Median	3409	3988	4056	3409
MIN	3413	3988	4056	3413
MAX	3412	3988	4056	3412
AUC	3411	3988	4056	3411
Mm	3413	3988	4056	3413
Rate	3412	3988	4056	3412
SD	3413	3988	4056	3413
CV	3412	3988	4056	3412
Total—Min	3403	3987	4055	3403

Table 2.40 Case 4 Model 21 (without interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	10.71	0.49	33.46	0.15	48.20	0.10	10.71
Mean	9.13	0.56	31.65	0.20	47.71	0.11	9.13
LO	10.36	0.51	33.18	0.16	48.12	0.10	10.36
FO	10.20	0.51	33.12	0.16	48.12	0.10	10.2
Median	9.65	0.54	32.25	0.18	47.86	0.11	9.65
MIN	10.14	0.52	33.01	0.16	48.09	0.10	10.14
MAX	10.29	0.51	32.99	0.17	48.08	0.10	10.29
AUC	9.1	0.57	31.66	0.2	47.72	0.11	9.1
Mm	9.91	0.53	32.67	0.17	48.01	0.11	9.91
Rate	10.57	0.5	33.21	0.16	48.13	0.1	10.57
SD	10.69	0.49	33.38	0.16	48.17	0.1	10.69
CV	10.56	0.5	33.23	0.16	48.14	0.1	10.56
Total—Min	9.1		31.65		47.71		9.1

Table 2.41 Case 4 Model 22 (longitudinal without interaction) CS)

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3092	3804	4037	3091
Mean	3082	3802	4038	3082
LO	3077	3801	4038	3077
FO	3057	3791	4035	3057
Median	3075	3800	4037	3075
MIN	3081	3800	4037	3081
MAX	3033	3782	4034	3033
AUC	3068	3796	4036	3068
Mm	3089	3802	4038	3089
Rate	3092	3804	4038	3092
SD	3089	3803	4038	3089
CV	3056	3791	4035	3056
Total—Min	3033	3782	4033	3033

Table 2.42 Case 4 Model 23 AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	14.58	0.08	14.81	0.11	9.26	0.23	9.26
Mean	14.35	0.10	14.79	0.11	9.25	0.23	9.25
LO	13.16	0.17	14.60	0.13	9.25	0.23	9.25
FO	14.57	0.08	14.82	0.11	9.26	0.23	9.26
Median	14.48	0.09	14.80	0.11	9.26	0.23	9.26
MIN	14.03	0.12	14.77	0.12	9.25	0.23	9.25
MAX	14.55	0.08	14.81	0.11	9.26	0.23	9.26
AUC	14.4	0.09	14.8	0.11	9.26	0.23	9.26
Mm	14.28	0.1	14.8	0.11	9.26	0.23	9.26
Rate	14.52	0.09	14.81	0.11	9.26	0.23	9.26
SD	14.46	0.09	14.8	0.11	9.26	0.23	9.26
CV	14.53	0.08	14.81	0.11	9.26	0.23	9.26
Total—Min	13.16		14.6		9.25		9.25

Table 2.43 Case 4 Model 23 CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	3.68	0.51	8.90	0.17	9.78	0.19	3.68
Mean	2.09	0.72	7.09	0.34	9.31	0.23	2.09
LO	3.34	0.56	8.65	0.19	9.72	0.19	3.34
FO	3.17	0.58	8.58	0.20	9.72	0.19	3.17
Median	2.62	0.65	7.71	0.28	9.46	0.22	2.62
MIN	3.12	0.59	8.48	0.21	9.69	0.20	3.12
MAX	3.27	0.57	8.46	0.21	9.68	0.20	3.27
AUC	2.06	0.73	7.1	0.34	9.32	0.23	2.06
Mm	2.88	0.62	8.13	0.24	9.61	0.2	2.88
Rate	3.55	0.53	8.68	0.19	9.73	0.19	3.55
SD	3.67	0.51	8.85	0.18	9.76	0.19	3.67
CV	3.55	0.53	8.7	0.19	9.74	0.19	3.55
Total—Min	2.06		7.09		9.31		2.06

Table 2.44 Case 4 Model 11 (with interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	24.13	0.11	46.56	0.07	49.22	0.10	24.13
Mean	23.79	0.12	46.47	0.07	49.20	0.10	23.79
LO	22.61	0.17	46.28	0.07	49.20	0.10	22.61
FO	24.01	0.11	46.50	0.07	49.21	0.10	24.01
Median	23.93	0.12	46.48	0.07	49.21	0.10	23.93
MIN	23.48	0.13	46.45	0.07	49.21	0.10	23.48
MAX	24.00	0.11	46.49	0.07	49.21	0.10	24
AUC	23.85	0.12	46.48	0.07	49.21	0.1	23.85
Mm	23.73	0.12	46.48	0.07	49.21	0.1	23.73
Rate	23.97	0.12	46.49	0.07	49.21	0.1	23.97
SD	23.91	0.12	46.48	0.07	49.21	0.1	23.91
CV	23.98	0.12	46.49	0.07	49.21	0.1	23.98
Total—Min	22.61		46.28		49.2		22.61

Table 2.45 Case 4 Model 12 (longitudinal with interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3387	3986	4054	3387
Mean	3386	3987	4055	3386
LO	3378	3986	4055	3378
FO	3388	3987	4055	3388
Median	3387	3987	4055	3387
MIN	3384	3987	4055	3384
MAX	3388	3987	4055	3388
AUC	3387	3987	4055	3387
Mm	3386	3987	4055	3386
Rate	3388	3987	4055	3388
SD	3387	3987	4055	3387
CV	3388	3987	4055	3388
Total—Min	3378	3986	4054	3378

Table 2.46 Case 4 Model 11 (with interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	9.42	0.55	32.75	0.17	47.70	0.11	9.42
Mean	7.84	0.63	30.92	0.22	47.21	0.12	7.84
LO	9.07	0.57	32.47	0.18	47.63	0.11	9.07
FO	8.91	0.58	32.41	0.18	47.62	0.11	8.91
Median	8.35	0.60	31.54	0.20	47.36	0.12	8.35
MIN	8.85	0.58	32.30	0.18	47.59	0.11	8.85
MAX	9.00	0.57	32.28	0.18	47.58	0.11	9
AUC	7.8	0.63	30.94	0.22	47.22	0.12	7.8
Mm	8.62	0.59	31.96	0.19	47.51	0.11	8.62
Rate	9.28	0.56	32.5	0.18	47.63	0.11	9.28
SD	9.4	0.55	32.67	0.17	47.67	0.11	9.4
CV	9.27	0.56	32.52	0.18	47.64	0.11	9.27
Total—Min	7.8		30.92		47.21		7.8

Table 2.47 Case 4 Model 12 (longitudinal with interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	2996	3792	4033	2996
Mean	2939	3770	4029	2939
LO	2987	3790	4033	2987
FO	2982	3790	4033	2982
Median	2962	3779	4030	2962
MIN	2980	3788	4033	2980
MAX	2985	3788	4033	2985
AUC	2938	3771	4029	2938
Mm	2972	3784	4032	2972
Rate	2994	3791	4033	2994
SD	2997	3793	4034	2997
CV	2994	3791	4033	2994
Total—Min	2938	3770	4029	2938

Table 2.48 Case 5 Model 21 (without interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	37.19	0.11	49.36	0.08	49.95	0.10	37.19
Mean	35.16	0.16	49.27	0.08	49.95	0.10	35.16
LO	36.36	0.13	49.23	0.08	49.94	0.10	36.36
FO	36.91	0.11	49.27	0.08	49.95	0.10	36.91
Median	36.15	0.13	49.29	0.08	49.95	0.10	36.15
MIN	36.92	0.11	49.29	0.08	49.95	0.10	36.92
MAX	32.82	0.21	49.18	0.08	49.94	0.10	32.82
AUC	34.85	0.16	49.27	0.08	49.95	0.1	34.85
Mm	34.25	0.18	49.24	0.08	49.95	0.1	34.25
Rate	35.27	0.15	49.23	0.08	49.94	0.1	35.27
SD	35.55	0.15	49.23	0.08	49.94	0.1	35.55
CV	35.32	0.15	49.22	0.08	49.94	0.1	35.32
Total—Min	32.82		49.18		49.94		32.82

Table 2.49 Case 5 Model 22 (longitudinal without interaction) AR(1)

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3720	4033	4058	3720
Mean	3714	4033	4059	3714
LO	3715	4033	4059	3715
FO	3711	4033	4059	3711
Median	3718	4033	4059	3718
MIN	3667	4032	4059	3667
MAX	3696	4033	4059	3696
AUC	3690	4032	4059	3690
Mm	3694	4032	4059	3694
Rate	3697	4032	4059	3697
SD	3694	4032	4059	3694
CV	3711	4033	4059	3711
Total—Min	3667	4032	4058	3667

Table 2.50 Case 5 Model 21 (without interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	28.19	0.22	41.41	0.13	49.58	0.10	28.19
Mean	23.59	0.35	40.04	0.16	49.49	0.10	23.59
LO	26.34	0.27	40.93	0.14	49.54	0.10	26.34
FO	27.91	0.23	41.26	0.13	49.56	0.10	27.91
Median	25.46	0.30	40.61	0.15	49.52	0.10	25.46
MIN	28.09	0.23	41.33	0.13	49.57	0.10	28.09
MAX	19.18	0.47	38.71	0.19	49.42	0.10	19.18
AUC	23.09	0.36	39.87	0.16	49.48	0.1	23.09
Mm	22.77	0.37	39.86	0.16	49.51	0.1	22.77
Rate	23.96	0.34	39.68	0.16	49.46	0.1	23.96
SD	24.86	0.31	40.06	0.16	49.48	0.1	24.86
CV	24.5	0.32	39.78	0.16	49.45	0.1	24.5
Total—Min	19.18		38.71		49.42		19.18

Table 2.51 Case 5 Model 22 (longitudinal without interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3531	3915	4052	3531
Mean	3506	3912	4053	3506
LO	3519	3914	4053	3519
FO	3503	3910	4053	3503
Median	3526	3915	4053	3526
MIN	3376	3888	4052	3376
MAX	3467	3903	4052	3467
AUC	3460	3902	4052	3460
Mm	3451	3896	4052	3451
Rate	3468	3900	4052	3468
SD	3462	3898	4052	3462
CV	3503	3910	4053	3503
Total—Min	3376	3888	4052	3376

Table 2.52 Case 5 Model 23 AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	25.36	0.04	15.68	0.11	9.37	0.23	9.37
Mean	23.49	0.11	15.66	0.11	9.38	0.23	9.38
LO	24.71	0.07	15.62	0.11	9.37	0.23	9.37
FO	25.28	0.05	15.66	0.11	9.38	0.23	9.38
Median	24.51	0.08	15.68	0.11	9.37	0.23	9.37
MIN	25.28	0.05	15.68	0.11	9.38	0.23	9.38
MAX	21.12	0.20	15.57	0.12	9.37	0.23	9.37
AUC	23.18	0.12	15.66	0.11	9.37	0.23	9.37
Mm	22.57	0.15	15.63	0.11	9.38	0.23	9.38
Rate	23.6	0.11	15.62	0.11	9.37	0.23	9.37
SD	23.9	0.1	15.62	0.11	9.37	0.23	9.37
CV	23.65	0.11	15.61	0.11	9.37	0.23	9.37
Total—Min	21.12		15.57		9.37		9.37

Table 2.53 Case 5 Model 23 CS

Covariance			AR(1)				
Correlation	0.9		0.5	0.1			
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	20.91	0.05	15.52	0.08	10.55	0.18	10.55
Mean	16.40	0.25	14.23	0.16	10.47	0.18	10.47
LO	19.19	0.13	15.12	0.11	10.52	0.18	10.52
FO	20.80	0.06	15.46	0.09	10.55	0.18	10.55
Median	18.30	0.17	14.80	0.13	10.50	0.18	10.5
MIN	20.98	0.05	15.53	0.08	10.55	0.18	10.55
MAX	11.90	0.45	12.87	0.24	10.41	0.19	10.41
AUC	15.88	0.27	14.05	0.17	10.46	0.19	10.46
Mm	15.56	0.29	14.04	0.17	10.49	0.18	10.49
Rate	16.77	0.24	13.86	0.18	10.44	0.19	10.44
SD	17.69	0.19	14.24	0.16	10.46	0.19	10.46
CV	17.32	0.21	13.96	0.17	10.43	0.19	10.43
Total—Min	11.9		12.87		10.41		10.41

Table 2.54 Case 5 Model 11 (with interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	36.51	0.12	48.95	0.09	49.55	0.11	36.51
Mean	34.45	0.17	48.86	0.09	49.54	0.11	34.45
LO	35.67	0.14	48.82	0.09	49.53	0.11	35.67
FO	36.23	0.13	48.86	0.09	49.54	0.11	36.23
Median	35.46	0.15	48.88	0.09	49.54	0.11	35.46
MIN	36.23	0.13	48.88	0.09	49.54	0.11	36.23
MAX	32.10	0.23	48.77	0.09	49.53	0.11	32.1
AUC	34.14	0.18	48.86	0.09	49.54	0.11	34.14
Mm	33.54	0.19	48.83	0.09	49.54	0.11	33.54
Rate	34.57	0.17	48.82	0.09	49.53	0.11	34.57
SD	34.85	0.16	48.82	0.09	49.53	0.11	34.85
CV	34.61	0.17	48.81	0.09	49.53	0.11	34.61
Total—Min	32.1		48.77		49.53		32.1

Table 2.55 Case 5 Model 12 (longitudinal with interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3704	4031	4058	3704
Mean	3683	4031	4059	3683
LO	3696	4031	4059	3696
FO	3698	4031	4059	3698
Median	3694	4032	4059	3694
MIN	3701	4032	4059	3701
MAX	3648	4031	4059	3648
AUC	3678	4031	4059	3678
Mm	3672	4031	4059	3672
Rate	3676	4031	4059	3676
SD	3679	4031	4059	3679
CV	3676	4031	4059	3676
Total—Min	3648	4031	4058	3648

Table 2.56 Case 5 Model 11 (with interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	26.80	0.26	40.56	0.15	49.01	0.11	26.8
Mean	22.16	0.39	39.19	0.18	48.91	0.11	22.16
LO	24.93	0.31	40.08	0.16	48.96	0.11	24.93
FO	26.52	0.27	40.42	0.15	48.99	0.11	26.52
Median	24.04	0.34	39.76	0.16	48.94	0.11	24.04
MIN	26.70	0.26	40.49	0.15	48.99	0.11	26.7
MAX	17.70	0.51	37.84	0.20	48.84	0.12	17.7
AUC	21.65	0.4	39.02	0.18	48.9	0.11	21.65
Mm	21.33	0.41	39	0.18	48.93	0.11	21.33
Rate	22.52	0.38	38.82	0.18	48.89	0.11	22.52
SD	23.44	0.35	39.2	0.17	48.9	0.11	23.44
CV	23.07	0.36	38.93	0.18	48.87	0.11	23.07
Total—Min	17.7		37.84		48.84	a	17.7

Table 2.57 Case 5 Model 12 (longitudinal with interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3477	3905	4050	3477
Mean	3418	3894	4050	3418
LO	3450	3901	4050	3450
FO	3463	3903	4050	3463
Median	3447	3899	4050	3447
MIN	3472	3905	4051	3472
MAX	3306	3877	4049	3306
AUC	3408	3892	4050	3408
Mm	3400	3891	4050	3400
Rate	3389	3885	4049	3389
SD	3407	3890	4050	3407
CV	3401	3887	4049	3401
Total—Min	3306	3877	4049	3306

Table 2.58 Case 6 Model 21 (without interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5	0.1			
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	86.27	0.40	86.09	0.35	76.43	0.34	76.43
Mean	82.98	0.42	86.00	0.35	76.46	0.34	76.46
LO	79.16	0.45	85.56	0.35	76.47	0.34	76.47
FO	85.67	0.40	86.10	0.35	76.47	0.34	76.47
Median	83.22	0.42	86.02	0.35	76.46	0.34	76.46
MIN	83.14	0.42	86.02	0.35	76.46	0.34	76.46
MAX	84.03	0.42	86.04	0.35	76.46	0.34	76.46
AUC	83.2	0.42	86.04	0.35	76.46	0.34	76.46
Mm	83.23	0.42	86.02	0.35	76.46	0.34	76.46
Rate	85.68	0.4	86.06	0.35	76.46	0.34	76.46
SD	85.51	0.41	86.03	0.35	76.47	0.34	76.47
CV	85.53	0.41	86.04	0.35	76.47	0.34	76.47
Total—Min	79.16		85.56		76.43		76.43

Table 2.59 Case 6 Model 22 (longitudinal without interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	4191	4270	4233	4191
Mean	4180	4269	4235	4180
LO	4190	4271	4235	4192
FO	4189	4271	4235	4189
Median	4190	4271	4235	4190
MIN	4191	4271	4235	4191
MAX	4190	4271	4235	4190
AUC	4190	4271	4235	4190
Mm	4192	4271	4235	4192
Rate	4192	4271	4235	4192
SD	4191	4271	4235	4191
CV	4189	4271	4235	4189
Total—Min	4180	4269	4233	4180

Table 2.60 Case 6 Model 21 (without interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	79.76	0.43	86.95	0.38	82.42	0.34	79.76
Mean	63.43	0.56	78.22	0.45	81.37	0.34	63.43
LO	67.49	0.53	83.62	0.41	82.26	0.34	67.49
FO	67.47	0.53	83.62	0.41	82.24	0.34	67.47
Median	64.76	0.55	79.92	0.43	81.74	0.34	64.76
MIN	66.69	0.53	82.47	0.41	82.07	0.34	66.69
MAX	66.64	0.53	82.66	0.41	82.18	0.34	66.64
AUC	63.4	0.56	78.29	0.45	81.4	0.34	63.4
Mm	64.9	0.55	80.65	0.43	81.88	0.34	64.9
Rate	77.68	0.45	86.74	0.38	82.44	0.34	77.68
SD	79.66	0.44	86.83	0.38	82.39	0.34	79.66
CV	76.95	0.46	86.67	0.38	82.44	0.34	76.95
Total—Min	63.4		78.22		81.37		63.4

Table 2.61 Case 6 Model 22 (longitudinal without interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null	3372	3951	4075	3372
Mean	2981	3853	4064	2981
LO	3097	3910	4072	3097
FO	3098	3909	4072	3098
Median	2991	3861	4065	2991
MIN	3034	3885	4069	3034
MAX	3034	3884	4069	3034
AUC	2979	3854	4064	2979
Mm	3000	3867	4066	3000
Rate	3243	3933	4073	3243
SD	3246	3934	4073	3246
CV	3242	3933	4073	3242
Total—Min	2979	3853	4064	2979

Table 2.62 Case 6 Model 23 AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	21.65	0.25	15.21	0.31	9.24	0.46	9.24
Mean	18.39	0.36	15.12	0.32	9.24	0.46	9.24
LO	14.50	0.49	14.67	0.34	9.24	0.46	9.24
FO	21.12	0.26	15.22	0.31	9.24	0.46	9.24
Median	18.63	0.35	15.14	0.32	9.24	0.46	9.24
MIN	18.54	0.35	15.14	0.32	9.24	0.46	9.24
MAX	19.45	0.32	15.16	0.31	9.24	0.46	9.24
AUC	18.61	0.35	15.16	0.31	9.24	0.46	9.24
Mm	18.63	0.35	15.14	0.32	9.24	0.46	9.24
Rate	21.12	0.26	15.18	0.31	9.24	0.46	9.24
SD	20.96	0.27	15.15	0.31	9.25	0.46	9.25
CV	20.98	0.27	15.16	0.31	9.24	0.46	9.24
Total—Min	14.5		14.67		9.24		9.24

Table 2.63 Case 6 Model 23 CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	18.79	0.27	17.10	0.31	11.06	0.42	11.06
Mean	2.24	0.91	8.26	0.66	9.97	0.48	2.24
LO	6.38	0.74	13.75	0.44	10.87	0.43	6.38
FO	6.36	0.74	13.74	0.44	10.85	0.43	6.36
Median	3.60	0.85	9.99	0.59	10.34	0.46	3.6
MIN	5.55	0.78	12.58	0.49	10.68	0.44	5.55
MAX	5.51	0.78	12.77	0.48	10.79	0.43	5.51
AUC	2.21	0.91	8.34	0.66	10	0.47	2.21
Mm	3.74	0.85	10.73	0.56	10.49	0.45	3.74
Rate	16.75	0.35	16.91	0.32	11.05	0.42	11.05
SD	18.75	0.28	17	0.31	11	0.42	11
CV	16	0.38	16.85	0.32	11.05	0.42	11.05
Total—Min	2.21		8.26		9.97		2.21

Table 2.64 Case 6 Model 11 (with interaction) AR(1)

Covariance			AR(1)				
Correlation	0.9		0.5	0.1			
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	33.05	0.76	48.78	0.62	49.20	0.56	33.05
Mean	29.65	0.78	48.63	0.62	49.18	0.56	29.65
LO	25.79	0.81	48.18	0.62	49.19	0.56	25.79
FO	32.36	0.76	48.72	0.62	49.19	0.56	32.36
Median	29.89	0.78	48.64	0.62	49.19	0.56	29.89
MIN	29.80	0.78	48.64	0.62	49.19	0.56	29.8
MAX	30.71	0.78	48.66	0.62	49.19	0.56	30.71
AUC	29.87	0.78	48.66	0.62	49.18	0.56	29.87
Mm	29.89	0.78	48.64	0.62	49.18	0.56	29.89
Rate	32.36	0.76	48.68	0.62	49.18	0.56	32.36
SD	32.2	0.77	48.65	0.62	49.19	0.56	32.2
CV	32.22	0.77	48.66	0.62	49.19	0.56	32.22
Total—Min	25.79		48.18		49.18		25.79

Table 2.65 Case 6 Model 12 (longitudinal with interaction) AR(1)

Covariance			AR(1)	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3511	4014	4050	3511
Mean	3495	4014	4050	3495
LO	3472	4011	4050	3472
FO	3509	4015	4050	3509
Median	3497	4014	4050	3497
MIN	3496	4014	4050	3496
MAX	3501	4014	4050	3501
AUC	3497	4014	4050	3497
Mm	3497	4014	4050	3497
Rate	3509	4014	4050	3509
SD	3508	4014	4051	3508
CV	3508	4014	4051	3508
Total—Min	3472	4011	4050	3472

Table 2.66 Case 6 Model 11 (with interaction) CS

Covariance			CS				
Correlation	0.9		0.5		0.1		
Covariates	MSE	R.adj	MSE	R.adj	MSE	R.adj	MIN-MSE
Null	24.86	0.81	43.95	0.68	51.24	0.58	24.86
Mean	8.30	0.94	35.07	0.74	50.14	0.59	8.3
LO	12.40	0.91	40.52	0.70	51.03	0.58	12.4
FO	12.38	0.91	40.51	0.70	51.01	0.58	12.38
Median	9.65	0.93	36.78	0.73	50.51	0.58	9.65
MIN	11.59	0.91	39.36	0.71	50.84	0.58	11.59
MAX	11.54	0.91	39.54	0.71	50.95	0.58	11.54
AUC	8.27	0.94	35.14	0.74	50.17	0.59	8.27
Mm	9.78	0.93	37.52	0.73	50.65	0.58	9.78
Rate	22.68	0.83	43.66	0.68	51.21	0.58	22.68
SD	24.67	0.81	43.75	0.68	51.16	0.58	24.67
CV	21.94	0.83	43.59	0.68	51.21	0.58	21.94
Total—Min	8.27		35.07		50.14		8.27

Table 2.67 Case 6 Model 12 (longitudinal with interaction) CS

Covariance			CS	
Correlation	0.9	0.5	0.1	MIN-AIC
Null—with—interaction	3186	3915	4071	3186
Mean	2973	3844	4061	2973
LO	3080	3895	4070	3080
FO	3080	3895	4070	3080
Median	3022	3863	4065	3022
MIN	3066	3886	4068	3066
MAX	3065	3887	4069	3065
AUC	2972	3844	4062	2972
Mm	3026	3870	4066	3026
Rate	3176	3915	4072	3176
SD	3187	3916	4071	3187
CV	3171	3915	4072	3171
Total—Min	2972	3844	4061	2972

References

ACCORD Study Group. (2010). Effects of intensive blood-pressure control in type 2 diabetes mellitus. *The New England journal of medicine, 362*(17), 1575.

Appel, L. J., Moore, T. J., Obarzanek, E., Vollmer, W. M., Svetkey, L. P., Sacks, F. M., … Lin, P. H. (1997). A clinical trial of the effects of dietary patterns on blood pressure. *New England Journal of Medicine, 336*(16), 1117–1124.

Appel, L. J., Wright Jr, J. T., Greene, T., Agodoa, L. Y., Astor, B. C., Bakris, G. L., … Gabbai, F. B. (2010). Intensive blood-pressure control in hypertensive chronic kidney disease. *New England Journal of Medicine, 363*(10), 918–929.

Assmann, S. F., Pocock, S. J., Enos, L. E., & Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet, 355*(9209), 1064–1069.

Bakx, J. C., van den Hoogen, H. J., van den Bosch, W. J., Van Schayck, C. P., van Ree, J. W., Thien, T., et al. (1999). Development of blood pressure and the incidence of hypertension in men and women over an 18-year period: Results of the Nijmegen Cohort Study. *Journal of Clinical Epidemiology, 52*(6), 531–538.

Blood Pressure Lowering Treatment Trialists' Collaboration. (2003). Effects of different blood-pressure-lowering regimens on major cardiovascular events: Results of prospectively-designed overviews of randomised trials. *The Lancet 362*(9395), 1527–1535.

Bristol, D. R. (2007). The choice of two baselines. *Drug information Journal, 41*(1), 57–61.

Chen, D. G., & Pounds, J. G. (1998). A non-linear isobologram model with Box-Cox transformation to both sides for chemical mixtures. *Environmental Health Perspectives, 106*(Suppl 6), 1367.

Clinical trials. https: //en.wikipedia.org/wiki/Clinical trial.

Ford, I., & Norrie, J. (2002). The role of covariates in estimating treatment effects and risk in longterm clinical trials. *Statistics in Medicine, 21*(19), 2899–2908.

Hansson, L., Zanchetti, A., Carruthers, S. G., Dahlf, B., Elmfeldt, D., Julius, S., … HOT Study Group. (1998). Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. *The Lancet, 351*(9118), 1755–1762.

Kramoh, E. K., Ngoran, Y. N., Ak-Traboulsi, E., Anzouan-Kacou, J. B., Konin, C. K., Coulibaly, I., … Guikahue, M. K. (2011). Hypertension management in an out patient clinic at the Institute of Cardiology of Abidjan (Ivory Coast). *Archives of Cardiovascular Diseases, 104*(11), 558–564.

Liu, G. F., Lu, K., Mogg, R., Mallick, M., & Mehrotra, D. V. (2009). Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? *Statistics in Medicine, 28*(20), 2509–2530.

Mizuno, J., & Monteiro, H. L. (2013). An assessment of a sequence of yoga exercises to patients with arterial hypertension. *Journal of Bodywork and Movement Therapies, 17*(1), 35–41.

Nussbaumerov, B., & Rosolov, H. (2013). Current opinion on aspirin in primary prevention of atherosclerotic cardiovascular diseases. Is there any difference between diabetic and non-diabetic patients? *cor et vasa, 55*(2), e190–e195.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine, 21*(19), 2917–2930.

Rdzanek, A., Filipiak, K. J., Karpiski, G., Grabowski, M., & Opolski, G. (2006). Exercise urotensin II dynamics in myocardial infarction survivors with and without hypertension. *International Journal of Cardiology, 110*(2), 175–178.

The Evolution of the Clinical Trials Process-A Brief History Lesson. http://www.psoriasiscouncil.org/docs/chapter01.pdf.

Tu, D., Shalay, K., & Pater, J. (2000). Adjustment of treatment effect for covariates in clinical trials: Statistical and regulatory issues. *Drug Information Journal, 34*(2), 511–523.

Ukena, C., Mahfoud, F., Kindermann, I., Barth, C., Lenski, M., Kindermann, M., … Sobotka, P. A. (2011). Cardiorespiratory response to exercise after renal sympathetic denervation in patients with resistant hypertension. *Journal of the American College of Cardiology, 58*(11), 1176–1182.

Van der Lee, J. H., Wagenaar, R. C., Lankhorst, G. J., Vogelaar, T. W., Devill, W. L., & Bouter, L. M. (1999). Forced use of the upper extremity in chronic stroke patients results from a single-blind randomized clinical trial. *Stroke, 30*(11), 2369–2375.

Weber, M. A., Jamerson, K., Bakris, G. L., Weir, M. R., Zappe, D., Zhang, Y., … Pitt, B. (2013). Effects of body size and hypertension treatments on cardiovascular event rates: Subanalysis of the ACCOMPLISH randomised controlled trial. *The Lancet, 381*(9866), 537–545.

Winkens, B., van Breukelen, G. J., Schouten, H. J., & Berger, M. P. (2007). Randomized clinical trials with a pre-and a post-treatment measurement: Repeated measures versus ANCOVA models. *Contemporary clinical trials, 28*(6), 713–719.

Zhang, P. G., Chen, D. G., & Roe, T. (2010). Choice of baselines in clinical trials: A simulation study from statistical power perspective. *Communications in Statistics Simulation and Computation, 39*(7), 1305–1317.

Chapter 3
Adaptive Trial Design in Clinical Research

Annpey Pong and Shein-Chung Chow

3.1 Introduction

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. Woodcock (2005) indicated that the low success rate of pharmaceutical development could be due to (i) a diminished margin for improvement that escalates the level of difficulty in proving drug benefits, (ii) genomics and other new science have not yet reached their full potential, (iii) mergers and other business arrangements have decreased candidates, (iv) easy targets are the focus as chronic diseases are harder to study, (v) failure rates have not improved, (vi) rapidly escalating costs and complexity decreases willingness/ability to bring many candidates forward into the clinic. 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 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 innovative adaptive design methods in clinical trials, while some researchers believe it is for the use of Bayesian approach.

The purpose of adaptive design methods in clinical trials is to give the investigator the flexibility for identifying best (optimal) clinical benefit of the test treatment under study without undermining the validity and integrity of the intended study. The concept of adaptive design can be traced back to 1970s when the adaptive

A. Pong (✉)
Merck Research Laboratories, Rahway, NJ, USA
e-mail: annpey.pong@merck.com

S.-C. Chow
Duke University School of Medicine, Durham, NC, USA
e-mail: sheinchung.chow@duke.edu

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_3

randomization and a class of designs for sequential clinical trials were introduced (Wei 1978). As a result, most adaptive design methods in clinical research are referred to as adaptive randomization (Wei 1978; Efron 1971; Lachin 1988; Atkinson and Donev 1992; Rosenberger et al. 2001; Rosenberger and Lachin 2002; Hardwick and Stout 2002), group sequential designs with the flexibility for stopping a trial early due to safety, futility and/or efficacy (Lan and DeMets 1987; Wang and Tsiatis 1987; Posch and Bauer 1999; Lehmacher and Wassmer 1999; Liu et al. 2002; Chow and Liu 2003), and sample size re-estimation at interim for achieving the desired statistical power (Cui et al. 1999; Shih 2001; Chung-Stein et al. 2006). The use of adaptive design methods for modifying the trial and/or statistical procedures of ongoing clinical trials based on accrued data has been practiced for years in clinical research. Adaptive design methods in clinical research are very attractive to clinical scientists due to the following reasons. First, it reflects medical practice in real world. Second, it is ethical with respect to both efficacy and safety (toxicity) of the test treatment under investigation. Third, it is not only flexible, but also efficient in the early phase of clinical development. 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 (EMEA 2002, 2006).

In recent years, the potential use of adaptive design methods in clinical trials have attracted much attention. The Pharmaceutical Research and Manufacturers of America (PhRMA) and Biotechnology Industry Organization (BIO) have established adaptive design working groups and proposed strategies, methodologies, and implementations for regulatory consideration (Gallo et al. 2006). However, there are no universal agreement in terms of definition, methodologies, and applications. Many journals have published special issues on adaptive design. These journals included, but are not limited to, Biometrics (Vol. 62, No. 3), Statistics in Medicine (Vol. 25, No. 19), Journal of Biopharmaceutical Statistics (Vol. 15, No. 4 and Vol. 17, No. 6), Biometrical Journal (Vol. 48, No. 4), and Pharmaceutical Statistics (Vol. 5, No. 2). For a comprehensive summarization of the issues and recommendations for the use of adaptive design methods, it may look up Chow and Chang (2006) and Chang (2007).

In the next section, commonly employed adaptations and the resultant adaptive designs are briefly described. Also included in this section are regulatory and statistical perspectives regarding the use of adaptive design methods in clinical trials. The impact of protocol amendments, challenges of *by design* adaptations, and obstacles of retrospective adaptations when applying adaptive design methods in clinical trials are discussed in Sects. 3.3–3.5, respectively. Some concluding remarks are given in the last section.

3.2 What Is Adaptive Design?

In clinical trials, it is not uncommon to modify trial procedures and/or statistical methods during the conduct of clinical trials based on the review of accrued data at interim. 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 the intended clinical trial. Trial procedures are referred to as the eligibility criteria, study dose, treatment duration, study endpoints, laboratory testing procedures, diagnostic procedures, criteria for evaluability, and assessment of clinical responses. Statistical methods include randomization scheme, study design selection, study objectives/hypotheses, sample size calculation, data monitoring and interim analysis, statistical analysis plan, and/or methods for data analysis. In this chapter, we will refer to the adaptations (changes or modifications) made to the trial and/or statistical procedures as the adaptive design methods. Thus, an adaptive design is defined as a design that allows adaptations to trial and/or statistical procedures of the trial after its initiation without undermining the validity and integrity of the trial (Chow et al. 2005). In one of their publications, with the emphasis of the feature of by design adaptations only (rather than ad hoc adaptations), the PhRMA Working Group on Adaptive Design refers to an adaptive design as a clinical trial design that uses accumulating data to decide on how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial (Gallo et al. 2006). In February 2010, a draft guidance on *Adaptive Design Clinical Trials for Drugs and Biologics* by the FDA was circulated for comments. The FDA draft guidance defines an adaptive design as a study that includes a *prospectively* planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of (usually interim) data from subjects in the study. The FDA draft guidance is a document describing the potential use of adaptive designs in clinical trials. It is generally viewed as supportive of the use of adaptive designs if they are employed properly. The FDA draft guidance is not a specific guidance for the implementation of adaptive designs in clinical trials. It, however, should be noted that adaptive designs have been used at times in confirmatory contexts, for the most part cautiously, limited to changes such as sample size re-estimation and treatment arm consolidation in the early phase of clinical development where there is more uncertainty and regulatory concerns are minimized. In many cases, an adaptive design is also known as a flexible design (EMEA 2002, 2006). Note that the 2010 FDA draft guidance on adaptive design is currently being revised by the FDA Adaptive Design Working Group.

3.2.1 Adaptations

An adaptation is referred to as a modification or a change made to trial procedures and/or statistical methods during the conduct of a clinical trial. By definition, adaptations that are commonly employed in clinical trials can be classified into the cate-

gories of prospective adaptation, concurrent (or ad hoc) adaptation, and retrospective adaptation. Prospective adaptations include, but are not limited to, adaptive randomization, stopping a trial early due to safety, futility or efficacy at interim analysis, dropping the losers (or inferior treatment groups), sample size re-estimation, and etc. Thus, prospective adaptations are usually referred to as by design adaptations as described in the PhRMA white paper (Gallo et al. 2006). Concurrent adaptations are usually referred to as any ad hoc modifications or changes made as the trial continues. Concurrent adaptations include, but are not limited to, modifications in inclusion/exclusion criteria, evaluability criteria, dose/regimen and treatment duration, changes in hypotheses and/or study endpoints, and etc. Retrospective adaptations are usually referred to as modifications and/or changes made to statistical analysis plan prior to database lock or unblinding of treatment codes. In practice, prospective, ad hoc, and retrospective adaptations are implemented by study protocol, protocol amendments, and statistical analysis plan with regulatory reviewer's consensus, respectively.

3.2.2 Type of Adaptive Designs

Based on the adaptations employed, commonly considered adaptive designs in clinical trials include, but are not limited to: (i) an adaptive randomization design, (ii) a group sequential a sample size re-estimation (SSRE) design, (iv) a drop-the-losers (or pick-the-winners) design, (v) an adaptive dose finding design, (vi) a biomarker-adaptive design, (vii) an adaptive treatment-switching design, (viii) an adaptive-hypothesis design, (ix) an adaptive seamless (e.g., phase I/II or phase II/III) trial design, and (x) a multiple adaptive design. These adaptive designs are briefly described below.

Adaptive randomization design—An adaptive randomization design is a design that allows modification of randomization schedules based on varied and/or unequal probabilities of treatment assignment in order to increase the probability of success. Commonly applied adaptive randomization procedures include treatment-adaptive randomization (Efron 1971; Lachin 1988), covariate-adaptive randomization, and response-adaptive randomization (Rosenberger et al. 2001; Rosenberger and Lachin 2002; Hardwick and Stout 2002). For example, a simple randomization may cause an imbalance of prognostic factors among treatment groups (e.g., a greater proportion of "very ill" subjects are assigned to a particular treatment group) for small or moderate clinical trials. The baseline adaptive randomization will improve the chance of having balance among treatment groups with respect to the prognostic factors.

Although an adaptive randomization design could increase the probability of success, it may not be feasible for a large trial or a trial with relatively long treatment duration because the randomization of a given subject depends on the response of the previous subject. A large trial or a trial with a relatively long treatment duration utilizing adaptive randomization design will take a much longer time to complete.

Besides, randomization schedule may not be available prior to the conduct of the study. Moreover, statistical inference on treatment effect is often difficult to obtain if it is not impossible due to complicated probability structure as the result of adaptive randomization.

Group sequential design—A group sequential design is a design that allows for prematurely stopping a trial due to safety, futility/efficacy or both with options of additional adaptations based on results of interim analysis. Various stopping boundaries based on different boundary functions for controlling an overall type I error rate are available in the literature (Rosenberger et al. 2001; Lan and DeMets 1987; Wang and Tsiatis 1987; Chow and Chang 2006; Jennison and Turnbull 2000, 2005). In recent years, the concept of two-stage adaptive design has led to the development of the adaptive group sequential design (Posch and Bauer 1999; Lehmacher and Wassmer 1999; Liu et al. 2002; Cui et al. 1999).
Unlike a traditional single-stage clinical trial in which the data remain blinded until the conclusion of the study, for example, a two-stage allows adaptive designs that permit increasing the sample size at the end of stage I while still protecting the type I error rate. It should be noted that the standard methods for group sequential design may not be appropriate (i.e., it may not be able to control the overall type I error rate at the desired level of 5%) if there is a shift in target patient population due to additional adaptations or protocol amendments.

Sample size re-estimation design—A sample size re-estimation (or N-adjustable) design is referred to as an adaptive design that allows for sample size adjustment or re-estimation based on the observed data at interim. Sample size adjustment or re-estimation could be done in either a blinding or unblinding fashion based on the criteria of treatment effect-size, conditional power, and/or reproducibility probability (Posch and Bauer 1999; Cui et al. 1999; Shih 2001; Chung-Stein et al. 2006; Proschan and Hunsberger 1995; Liu and Chi 2001; Friede and Kieser 2004). Sample size re-estimation suffers from the same disadvantage as the original power analysis for sample size calculation prior to the conduct of the study because it is performed by treating estimates of the study parameters, which are obtained based on data observed at interim, as true values. It is not a good clinical/statistical practice to start with a small number and then perform sample size re-estimation (adjustment) at interim by ignoring the clinically meaningful difference that one wishes to detect for the intended clinical trial. It should be noted that the observed difference at interim based on a small number of subjects may not be of statistically significant (i.e., it may be observed by chance alone). In addition, there is variation associated with the observed difference which is an estimate of the true difference. Thus, standard methods for sample size re-estimation based on the observed difference with a limited number of subjects may be biased and misleading.

Drop-the-losers design—A drop-the-losers design is a design that allows dropping the inferior treatment groups. A drop-the-losers design may also allow adding additional arms. A drop-the-losers design is useful in phase II clinical development especially when there are uncertainties regarding the dose levels (Bauer and Kieser

1999; Brannath et al. 2003; Sampson and Sill 2005; Posch et al. 2005). The selection criteria and decision rules play important roles for drop-the-losers designs. Dose groups that are dropped may contain valuable information regarding dose response of the treatment under study. Typically, drop-the-losers design is a two-stage design. At the end of the first stage, the inferior arms will be dropped based on some pre-specified criteria. The winners will then proceed to the next stage. In practice, the study is often powered for achieving a desired power at the end of the second stage (or at the end of the study). In other words, there may not be any statistical power for the analysis at the end of the first stage for dropping the losers (or picking up the winners). In this case, it is a common practice to drop the losers or pick up the winners based on so-called precision analysis, i.e., an approach for determining the confidence level for achieving a statistical significance. Note that some clinical scientists prefer the term pick-the-winners designs.

Adaptive dose finding design—An adaptive dose finding (e.g., escalation) design is often used in early phase clinical development to identify the minimum effective dose (MED) and/or the maximum tolerable dose (MTD), which is used to determine the dose level for the next phase clinical trials (Bauer and Rohmel 1995; Whitehead 1997; Zhang et al. 2006). For adaptive dose finding design, the method of continual re-assessment method (CRM) in conjunction with Bayesian approach is usually considered (O'Quigley et al. 1990; O'Quigley and Shen 1996; Chang and Chow 2005). Mugno et al. (2004) introduced a nonparametric adaptive urn design approach for estimating a dose-response curve. More details regarding PhRMA's proposed statistical methods, the reader may consult with a special issue recently published at the Journal of Biopharmaceutical Statistics, Vol. 17, No. 6.

Biomarker-adaptive design—A biomarker-adaptive design is a design that allows for adaptations based on the response of biomarkers such as genomic markers. An adaptive-biomarker design involves biomarker qualification and standard, optimal screening design, and model selection and validation. It should be noted that there is a gap between identifying biomarkers that associated with clinical outcomes and establishing a predictive model between relevant biomarkers and clinical outcomes in clinical development. For example, correlation between biomarker and true clinical endpoint makes a prognostic marker. However, correlation between biomarker and true clinical endpoint does not make a predictive biomarker. 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. A *predictive biomarker* informs the treatment effect on the clinical endpoint (Chang 2007).
A biomarker-adaptive design can be used to (i) select right patient population (e.g., enrichment process for selection of a better target patient population), (ii) identify nature course of disease, (iii) early detection of disease, and (iv) help in developing personalized medicine (Chang 2007; Charkravarty 2005; Wang et al. 2007).

Adaptive treatment-switching design—An adaptive treatment-switching design is a design that allows the investigator to switch a patient's treatment from an initial assignment to an alternative treatment if there is evidence of lack of efficacy or safety of the initial treatment (Shao et al. 2005; Branson and Whitehead 2002). In cancer clinical trials, estimation of survival is a challenge when treatment-switching has occurred in some patients. A high percentage of subjects who switched due to disease progression could lead to change in hypotheses to be tested. In this case, sample size adjustment for achieving a desired power is necessary.

Adaptive-hypotheses design—An adaptive-hypotheses design refers to a design that allows modifications or changes in hypotheses based on interim analysis results (Hommel 2001). Adaptive-hypotheses designs often considered before database lock and/or prior to data unblinding. Some examples include the switch from a superiority hypothesis to a non-inferiority hypothesis and the switch between the primary study endpoint and the secondary endpoints. For the switch from a superiority hypothesis to a non-inferiority hypothesis, the selection of non-inferiority margin is critical which has an impact on sample size adjustment for achieving the desired power. According to the ICH guideline, the selected non-inferiority margin should be both clinical and statistical justifiable (International Conference on Harmonization Guideline E10 2000; Chow and Shao 2005).

Adaptive seamless trial design—An adaptive seamless trial design is referred to a program that addresses within single trial objectives that are normally achieved through separate trials of clinical development. An adaptive seamless design is an adaptive seamless trial design that would use data from patients enrolled before and after the adaptation in the final analysis (Kelly et al. 2005; Maca et al. 2006). Commonly considered adaptive seamless trials in clinical development are an adaptive seamless phase I/II design in early clinical development and an adaptive seamless phase II/III trial design in late phase clinical development. An adaptive seamless phase II/III design is a two-stage design consisting of a learning or exploratory stage (phase IIb) and a confirmatory stage (phase III). A typical approach is to power the study for the phase III confirmatory phase and obtain valuable information with certain assurance using confidence interval approach at the phase II learning stage. Its validity and efficiency, however, has been challenged (Tsiatis and Mehta 2003). Moreover, it is not clear how to perform a combined analysis if the study objectives (or endpoints) are similar but different at different phases. (Chow et al. 2007) Further research is needed.

Multiple adaptive design—Finally, a multiple adaptive design is any combinations of the above adaptive designs. Commonly considered multiple-adaptation designs include (i) the combination of adaptive group sequential design, drop-the-losers design, and adaptive seamless trial design and (ii) adaptive dose-escalation design with adaptive randomization (Chow and Chang 2006). In practice, since statistical inference for a multiple-adaptation design is often difficult, it is suggested that a clinical trial simulation be conducted to evaluate the performance of the resultant multiple adaptive design at the planning stage.

Note that as indicated in its draft guidance, the FDA classifies adaptive designs into *well-understood* designs and *less well-understood* designs. Well-understood design refers to the typical group sequential design, which has been employed in clinical research for years. Less well-understood designs include the adaptive dose finding and two-stage phase I/II (or II/III) seamless designs. Many scientific issues surrounding the less well-understood designs are posted in the draft guidance without recommendations for resolution. This raises the question whether the use of adaptive design methods in clinical trials (especially for those less well-understood designs) is ready for implementation in practice.

3.2.3 Regulatory/Statistical Perspectives

From regulatory point of view, the use of adaptive design methods based on accrued data in clinical trials may introduce operational bias such as selection bias, method of evaluation, early withdrawal, and modification of treatment. Consequently, it may not be able to preserve the overall type I error rate at the pre-specified level of significance. In addition, p-values may not be correct and the corresponding confidence intervals for the treatment effect may not be reliable. Moreover, it may result in a totally different trial that is unable to address the medical questions that original study intended to answer. Li (2006) also indicated that commonly seen adaptations which have an impact on the type I error rate include, but are not limited to, (i) sample size adjustment at interim, (ii) sample size allocation to treatments, (iii) delete, add, or change treatment arms, (iv) shift in target patient population such as changes in inclusion/exclusion criteria, (v) change in statistical test strategy, (vi) change in study endpoints, and (vii) change in study objectives such as the switch from a superiority trial to a non-inferiority trial. As a result, it is difficult to interpret the clinically meaningful effect size for the treatments under study (Quinlan et al. 2006).

From statistical point of view, major (or significant) adaptations to trial and/or statistical procedures could (i) introduce bias/variation to data collection, (ii) result in a shift in location and scale of the target patient population, and (iii) lead to inconsistency between hypotheses to be tested and the corresponding statistical tests. These concerns will not only have an impact on the accuracy and reliability of statistical inference drawn on the treatment effect, but also present challenges to biostatisticians for development of appropriate statistical methodology for an unbiased and fair assessment of the treatment effect.

Note that although the flexibility of modifying study parameters is very attractive to clinical scientists, several regulatory questions/concerns arise. First, what level of modifications to the trial procedures and/or statistical procedures would be acceptable to the regulatory authorities? Second, what are the regulatory requirements and standards for review and approval process of clinical data obtained from adaptive clinical trials with different levels of modifications to trial procedures and/or statistical procedures of ongoing clinical trials? Third, has the clinical trial become a totally different clinical trial after the modifications to the trial procedures and/or statistical

procedures for addressing the study objectives of the originally planned clinical trial? These concerns should be addressed by the regulatory authorities before the adaptive design methods can be widely accepted in clinical research and development.

3.3 Impact of Protocol Amendments

3.3.1 Moving Target Patient Population

In practice, for a given clinical trial, it is not uncommon to have 3–5 protocol amendments after the initiation of the clinical trial. One of the major impacts of many protocol amendments is that the target patient population may have been shifted during the process, which may have resulted in a totally different target patient population at the end of the trial. A typical example is the case when significant adaptation (modification) is applied to inclusion/exclusion criteria of the study. Denote by the target patient population with the mean μ and standard deviation σ for the primary endpoint. After a given protocol amendment, the resultant (actual) patient population may have been shifted to (μ_1, σ_1), where $\mu_1 = \mu + \varepsilon$ is the population mean of the primary study endpoint and $\sigma_1 = C\sigma$ $(C > 0)$ is the population standard deviation of the primary study endpoint. The shift in target patient population can be characterized by $\left|\frac{\mu_1}{\sigma_1}\right| = \left|\frac{\mu+\varepsilon}{C\sigma}\right| = |\Delta|\left|\frac{\mu}{\sigma}\right|$, where $\Delta = \frac{1+\varepsilon/\mu}{C}$. Chow and Chang (2006) refer to Δ as a sensitivity index measuring the change in effect size between the actual patient population and the original target patient population.

Denote by, (μ_i, σ_i) the actual patient population after the *i*th modification of trial procedure, where $\mu_i = \mu + \varepsilon_i$ and $\sigma_i = C_i\sigma$, $i = 0, 1, \ldots K$. Note that $i = 0$ reduces to the original target patient population (μ, σ). That is, when $i = 0$, $\varepsilon_0 = 0$ and $C_0 = 1$. After K protocol amendments, the resultant actual patient population becomes (μ_K, σ_K), where $\mu_k = \mu + \sum_{i=1}^{K} \varepsilon_i$ and $\sigma_k = \sum_{i=1}^{K} C_i\sigma$. It should be noted that (ε_i, C_i), $i = 1,\ldots, K$ are in fact random variables. As a result, the resultant actual patient population following certain modifications to the trial procedures is a *moving* target patient population rather than a fixed target patient population.

The impact of protocol amendments on statistical inference due to shift in target patient population (moving target patient population) can be studied through a model that link the moving population means with some covariates (Chow and Chang 2006; Chow and Shao 2006). Chow and Shao (2005) derived statistical inference for the original target patient population for simple cases. Their approach and recommendations for improvement are briefly outline below (Chow and Shao 2006).

3.3.2 Statistical Inference with Covariate Adjustment

Suppose that there are a total of K protocol amendments for a given clinical trial that compares I treatments. Let μ_{i0} be the mean of the study endpoint of the original target patient population under treatment i and μ_{ik} be the mean of the patient population under treatment i after the kth protocol amendment. Suppose that the parameters of interest are μ_{i0}'s (not μ_{ik}) with $k = 1,\ldots, K$. If the differences among μ_{i0}, $\mu_{i1},\ldots,$ μ_{ik} are ignored and statistical inference is made by pooling all data (before and after protocol amendment), then the conclusion drawn on μ_{i0}'s may be biased and misleading. Assuming that there is a relationship between μ_{ik}'s and a covariate vector x. Chow and Shao (2005) considered the following regression model (Chow and Shao 2006).

$$\mu_{ik} = \alpha_i + \beta_i' x_{ik} \quad k = 0, 1, \ldots, K, \quad i = 1, \ldots, I, \tag{3.1}$$

where α_i is an unknown parameter, β_i is a vector of unknown parameters, and β_i' is the transpose of β_i. The best example is the change of patient inclusion/exclusion criterion due to the problem of not enough patients. In Model (3.1), it is assumed x_{ik} is fixed and known so that α_i and β_i can be estimated through a regression between $\bar{y}_{ik}$ and $\bar{x}_{ik}$, where $\bar{y}_{ik}$ and $\bar{x}_{ik}$ are the sample mean of the study endpoint and the sample mean of the fixed known x-covariates, respectively, under treatment i after modification k, $i = 1,\ldots, I$, $k = 0,1,\ldots, K$. Once α_i and β_i are estimated, μ_{i0} can be estimated from $\mu_{i0} = \alpha_i + \beta_i' \bar{x}_{i0}$.

In practice, however, x_{ik} is often an observed random covariate vector and model (3.1) should be replaced by

$$\mu_{ik} = \alpha_i + \beta_i' v_{ik} \quad k = 0, 1, \ldots, K, \quad i = 1, \ldots, I, \tag{3.2}$$

where v_{ik} is the population mean of the covariate under treatment i after modification k. Let y_{ikj} be the observed study endpoint from the jth patient under treatment i after amendment k and x_{ikj} be the associated observed covariate, $j = 1,\ldots, n_{ik}$, $k = 0,1,\ldots,$ K, $i = 1,\ldots, I$. There is room for improvement by the following two directions.

First, under model (3.2), we estimate μ_{ik} by $\bar{y}_{ik}$ (the sample mean of $y_{ik_1},\ldots, y_{ikn_{ik}}$) and v_{ik} by $\bar{x}_{ik}$ (the sample mean of $x_{ik_1},\ldots, x_{ikn_{ik}}$). Then, we estimate α_i and β_i by the weighted least squares estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ in a "regression" between $\bar{y}_{ik}$ and $\bar{x}_{ik}$ for each fixed i. The parameter μ_{i0} is estimated by $\hat{\mu}_{i0} = \hat{\alpha}_i + \hat{\beta}_i' \bar{x}_{i0}$. Statistical inference (such as hypothesis testing and confidence intervals) can be made using $\hat{\mu}_{i0}$ and its exact or asymptotic distribution can be derived accordingly. Second, a more efficient statistical inference can be made if we replace model (3.2) by the following stronger model:

$$E(y_{ikj}|x_{ikj}) = \alpha_i + \beta_i' x_{ikj}, \quad {}_{j=1,\ldots,n_{ik},} \quad k = 0, 1, \ldots, K, \quad i = 1, \ldots, I. \tag{3.3}$$

Under this model, we can first fit a regression between y_{ikj} and x_{ikj} for each fixed i to obtain the least squares estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$. Then, μ_{i0} is estimated by $\hat{\mu}_{i0} = \hat{\alpha}_i + \hat{\beta}_i' \bar{x}_{i0}$. Statistical inference can be made using $\hat{\mu}_{i0}$ and its exact or asymptotic distribution can be derived. Note that model (3.3) is stronger than model (3.2) so that we need to balance the gain in efficiency over bias due to the violation of model (3.3).

As an alternative to model (3.1), we may consider a random-deviation model. Suppose that there exist random variables δ_{ik}, $k = 1,\ldots, K$, $i = 1,\ldots, I$, such that

$$\mu_{ik} = \mu_{i0} + \delta_{ik}, \quad k = 1, \ldots, K, i = 1, \ldots, I. \tag{3.4}$$

This means that the population mean after the kth protocol amendment deviates from the mean of the target population by a random effect δ_{ik}. Of course, we may consider combining models (3.1) and (3.4):

$$\mu_{ik} = \alpha_i + \beta_i' x_{ik} + \delta_{ik}, \quad k = 1, \ldots, K, \quad i = 1, \ldots I.$$

Under model (3.4) and the assumptions that, conditional on δ_{ik}'s, y_{ikj}'s are independent with mean μ_{ik} in (3.4) and variance σ^2, where y_{ikj} is the study endpoint for the jth patient under treatment i after modification k. Then, the observed data follow a mixed effects model. Consequently, the existing statistical procedures for mixed effects models can be applied to the estimation or inference. A further assumption can be imposed to model (3.4).

$$y_{ikj} = \mu_{i0} + \lambda_k + \gamma_{ik} + \varepsilon_{ikj}, \quad i = 1,\ldots,I, \quad \mathrm{k} = 1,\ldots,K, \tag{3.5}$$

where λ_k's, γ_{ik}'s (which reflect the "interaction" between treatment and λ_k), ε_{ikj} are independently normal distributed with mean zero and variances σ_λ^2, σ_γ^2, and σ^2, respectively. Gallo and Khuri (1990) derived an exact test for the unbalanced mixed effects model (3.5). Although other refinements were developed (Ofversten 1993; Christensen 1996; Khuri et al. 1998), these existing tests do not have explicit forms; so the computation is complicated.

Not that model (3.4) can be modified under the Bayesian frame work. For future methodology development based on statistical inference with covariate adjustment, it is of interest to consider (i) a more complex situations such as unequal sample sizes n_{ik} and/or unequal variances after protocol amendments, (ii) deriving approximations to the integrals involved in the posterior probabilities, (iii) studying robustness of the choices of prior parameters, and (iv) alternative forms of null hypotheses such as $-a \leq \theta \leq a$ with a given $a > 0$.

3.3.3 *Inference Based on Mixture Distribution*

The primary assumption of the above approaches is that there is a relationship between μ_{ik}'s and a covariate vector x. In practice, such covariates may not exist or may not be observable. In this case, it is suggested assessing the sensitivity index and consequently deriving a unconditional inference for the original target patient population assuming that the shift parameter (i.e., ε) and/or the scale parameter (i.e., C) is random (Chow and Chang 2006). It should be noted that effect of ε_i could be offset by C_i for a given modification i as well as by (ε_j, C_j) for another modification j. As a result, estimates of the effects of (ε_i, C_i), $i = 1,\ldots, K$ are difficult, if not impossible, to obtain. In practice, it is desirable to limit the combined effects of (ε_i, C_i), $i = 0, \ldots, K$ to an acceptable range for a valid and unbiased assessment of treatment effect regarding the target patient population based on clinical data collected from the actual patient population.

3.4 Challenges in *by Design* Adaptations

In clinical trials, commonly employed by design (prospective) adaptations include stopping the trial early due to safety, futility, and/or efficacy, sample size re-estimation (adaptive group sequential design), dropping the losers (adaptive dose finding design), and combining two separate trials into a single trial (adaptive seamless design) . In this section, major challenges in these by design adaptations will be described. Recommendations for resolution are provided whenever possible.

3.4.1 *Adaptive Group Sequential Design*

The group sequential design has a long history of application in clinical trials. It is a design in which the accumulating data are analyzed at a series of interim analyses during the course of the trial. The main purpose of the group sequential design is to allow the trail to be stopped for clinical benefit or harm at an interim analysis. The idea of adaptive group sequential design started in early 1990s. It allows a wide range of modifications to the trial design at each interim analysis. For example, the sample size for a future interim analysis may be re-calculated based on the observed treatment differences at the current interim analysis using accumulated data. Results on sample size re-calculation and other modifications can be found in the literature (Lehmacher and Wassmer 1999; Cui et al. 1999; Jennison and Turnbull 2005; Proschan and Hunsberger 1995; Liu and Chi 2001; Shen and Fisher 1999; Posch and Bauer 2000; Hommel et al. 2005; Hung et al. 2005; Li et al. 2005; Proschan 2005; Kelly et al. 2005; Proschan et al. 2005).

Consider a group sequence design with K interim analyses without adaptation (i.e., there is no change in the trial design after each interim analysis). Let Z_k be the test statistic at the kth interim analysis such that H_0 is rejected and the trial will be stopped if $|Z_k| > c_{\alpha,k,\kappa}$, where $c_{\alpha,k,\kappa}$ is a constant depending on k, κ, and the significance level α. For $k = 1, \ldots, K-1$, the trial continues if $|Z_k| \le c_{\alpha,k,\kappa}$. At the last stage $k = K$, the trial stops and H_0 is concluded if $|Z_k| \le c_{\alpha,k,\kappa}$. Consequently, the overall type I error rate for this procedure is

$$\alpha_K = \sum_{k=1}^{K} P(|Z_k| > c_{\alpha,k,K}, |Z_1| \le c_{\alpha,1,K}, \ldots, |Z_{K-1}| \le c_{\alpha,k-1,K} \,|H_0). \quad (3.6)$$

If we choose $c_{\alpha,k,\kappa}$'s to satisfy $\alpha_K = \alpha$, then the overall type I error rate is maintained. The selection of $c_{\alpha,k,\kappa}$'s relies on the distributions of Z_k's under H_0. When Z_k is a standard normal random variable based on data accumulated up to the kth interim analysis, the most popular choice of $c_{\alpha,k,\kappa}$'s is Pocock's $c_{\alpha,1,\kappa} = \cdots = c_{\alpha,k,\kappa}$ given in a table by Pocock (1977). Since $c_{\alpha,k,\kappa}$'s are not uniquely determined, various modifications to Pocock's test have been proposed, for example, O'Brien and Fleming's test, Wang and Tsiatis' test, and the inner wedge test (Jennison and Turnbull 2000).

The assumption that Z_k's are standard normal, however, does not hold in most practical situations. It holds approximately when the sample size at each interim analysis is large enough, based on the central limit theorem. Consider for example a parallel-group design with two treatments. Let x_{ikj} be the response from patient j under treatment i at the kth interim analysis, $i = 1, 2$, $j = 1, \ldots, n$, $k = 1, \ldots, K$. Let $\bar{x}_{ik} = \frac{1}{kn}\sum_{j=1}^{n} x_{ikj}$, $i = 1, 2$, $k = 1, \ldots, K$. Assume x_{ikj}'s are independently normal with mean μ_i and variance σ^2. If H_0: $\mu_1 = \mu_2$ and σ^2 is known, then $Z_k = \sqrt{kn}(\bar{x}_{1k} - \bar{x}_{2k})/\sqrt{2\sigma^2}$ is standard normal under H_0, $k = 1, \ldots, K$, and α_K in (3.6) is exactly equal to α with Pocock's $c_{\alpha,k,\kappa}$. In practice, σ^2 is usually unknown and has to be estimated. If σ^2 is replaced by the pooled sample variance, then Z_k has a t-distribution instead of the standard normal. As a result, α_K in (3.6) is not α and the overall type I error rate is not maintained. Of course, $\alpha_K \to \alpha$ as $n \to \infty$, because of the central limit theorem.

One of major challenges for an adaptive group sequential design is that the overall type I error rate may be in inflated when there is a shift in target patient population (Feng et al. 2007). To overcome this problem, we suggest studying the effect of unknown σ^2 and sample size n in the selection of $c_{\alpha,k,\kappa}$, for various methods and study designs. With Pocock's method, for example, the simulation results in Table 3.1 show that the overall type I error rate α_K can be much larger than α (=0.05), when σ^2 is replaced by the pooled sample variance (i.e., each Z_k is not normal but t-distributed). The simulation was carried out using two treatments, equal sample size n for any treatment and interim analysis, up to $K = 5$ stages, and simulation size 50,000. For $K = 2, 3, 4, 5$, Pocock's $c_{0.05,k,K}$ values do not depend on k and are 2.178, 2.289, 3.361, 4.413, respectively.

From Table 3.1, the overall type I error rates are greater than the nominal level 0.05. As the sample size n increases, these error rates are closer to 0.05. But the

Table 3.1 Overall type I error rate for Pocock's test

n	K			
	2	3	4	5
2	0.197	0.198	0.195	0.198
3	0.129	0.127	0.127	0.124
4	0.100	0.101	0.101	0.100
5	0.089	0.089	0.086	0.086
6	0.081	0.079	0.079	0.079
7	0.076	0.074	0.075	0.075
8	0.071	0.070	0.071	0.071
9	0.069	0.069	0.068	0.069
10	0.067	0.067	0.065	0.067
11	0.066	0.063	0.066	0.066
12	0.063	0.065	0.065	0.064
13	0.061	0.059	0.061	0.061
14	0.063	0.062	0.062	0.061
15	0.062	0.061	0.061	0.060
16	0.061	0.058	0.059	0.059
17	0.058	0.059	0.059	0.058
18	0.060	0.056	0.057	0.060
19	0.058	0.059	0.058	0.058
20	0.058	0.059	0.056	0.058
21	0.059	0.058	0.056	0.056
22	0.057	0.057	0.057	0.057
23	0.058	0.055	0.056	0.057
24	0.057	0.057	0.055	0.058
25	0.056	0.056	0.058	0.054
26	0.057	0.055	0.057	0.056
27	0.056	0.055	0.055	0.053
28	0.056	0.055	0.055	0.055
29	0.054	0.056	0.055	0.055
30	0.054	0.057	0.054	0.055

results for small n are not satisfactory. We also note that the number of stages does not have a large effect, since the overall type I error rates are quite close across the stages when the sample size n is fixed. Thus, we suggest developing a new procedure of choosing $c_{\alpha,k,\kappa}$'s so that the overall type I error rate α_K is exactly equal to α, when σ^2 is replaced by the pooled sample variance in various group sequential tests. The new $c_{\alpha,k,\kappa}$ depends on the sample size n. A statistical table or software for $c_{\alpha,k,\kappa}$ with different n will be constructed.

An adaptive group sequential design is attractive to sponsors in clinical development because it allows adaptations of the trial after each interim to meet specific needs within limited budget or resources and target timelines. However, some adaptations

may introduce bias/variation to data collection as the trial continues. To account for these (expected and/or unexpected) biases/variation, statistical tests are necessary adjusted to maintain the overall type I error and the related sample size calculation formulas have to be modified for achieving the desired power. Statistical inference on moving patient population described in the previous section can be applied to the adaptive group sequential design when the adaptations at each interim analysis may alter the patient population. The problem for the adaptive group sequential design is more difficult, since the test procedure is much more complicated. Thus, for future development, it is worthy pursuing the following specific directions that (i) deriving valid statistical test procedures for adaptive group sequential designs assuming model (3.2) or (3.3), which relates the data from different interim analyses, (ii) deriving valid statistical test procedures for adaptive group sequential designs assuming the random-deviation model (3.5), (iii) deriving valid Bayesian methods for adaptive group sequential designs, and (iv) deriving sample size calculation formulas for various situations.

3.4.2 Adaptive Dose Finding Design

Chang and Chow (2005) proposed a hybrid Bayesian adaptive design for dose escalation study, which involves the steps of (i) construction of utility function, (ii) probability model for dose-response, (iii) the selection of a prior, (iv) the re-assessment of model parameters, (v) update of the utility function, (vi) the determination of the next action (i.e., treatment assignment of the next subject near the estimated MTD).

In early clinical development for establishing a dose response relationship, an adaptive multistage design is commonly used (Bauer and Rohmel 1995). The adaptive multistage design allows various adaptations of the design in one or two pre-scheduled interim analyses. The idea of an adaptive multistage design for establishing dose response relationship is not only to reassess the sample size by using the observed variability and/or effects, but also to reduce the set of multiple endpoints to suitable subsets. In addition, the adaptive multistage design allows selecting a subset of doses for further experimentation. More specifically, consider a two-stage scenario. One would start with two doses from the conjectured therapeutic dose range. If in the pre-scheduled interim analysis there is no sufficient trend, the doses may be changed (e.g., by lowering the low dose and/or increasing the high dose). The second stage of the experiment will be performed and the overall decision relies on a combination test of the test results from the two stages separately. It, however, should be noted that the adaptive multistage design for establishing dose response relationship suffers the disadvantage that only the p-values of the samples are combined but no pooling of the samples itself is performed. This leads to a crucial point of interpretation.

For dose-toxicity studies, the "3+3" or more generally, the "m+n" traditional escalation rules (TER) are commonly used in early phase of oncology studies. Many new methods such as the assessment of dose response using multiple-stage designs

(Crowley 2001) and the continued reassessment method (CRM) (O'Quigley et al. 1990; O'Quigley and Shen 1996; Babb and Rogatko 2001) have been developed. For the method of CRM, the dose-response relationship is continually reassessed based on accumulative data collected from the trial. The next patient who enters the trial is then assigned to the potential MTD level. This approach is more efficient than that of the usual TER with respect to the allocation of the MTD. However, the efficiency of CRM may be at risk due to delayed response and/or a constraint on dose-jump in practice (Babb and Rogatko 2001). In recent years, the use of adaptive design methods for characterizing dose response curve has become very popular (Bauer and Rohmel 1995). An adaptive design is a dynamic system that allows the investigator to optimize the trial (including design, monitoring, operating, and analysis) with cumulative information observed from the trial. For Bayesian adaptive design for dose response trials, some researchers suggest the use of loss/utility function in conjunction with dose assignment based on minimization/maximization of loss/utility function (Whitehead 1997; Gasprini and Eisele 2000).

Let $X = \{x_1, x_2, \ldots, x_K\}$ be the action space where x_i is the coded value for action of anything that would affect the outcomes or decision-making such as a treatment, a withdrawal of a treatment arm, a protocol amendment, stopping the trial or any combination of the above. In practice, x_i can be either a fixed dose or a variable dose given to a patient. If action x_i is not taken, then $x_i = 0$. Let $y = \{y_1, y_2, \ldots, y_m\}$ be the outcomes of interest, which can be efficacy or toxicity of a test treatment. In each of these outcomes, y_i is a function of action $y_i(x)$, $x \in X$. The utility is then defined as $U = \sum_{j=1}^{m} w_j = \sum_{j=1}^{m} w(y_j)$, where U is normalized such that $0 \leq U \leq 1$ and w_j are pre-specified weights.

To allow more patients to be assigned to superior treatment groups, the target randomization probability to x_i group should be proportional to the current estimation of utility or response rate of the group, that is $U(x_i)/\sum_{i=1}^{K} U(x_i)$ where K is the number of groups. As a result, the utility-adaptive randomization can be given. It is to be noted that the hybrid Bayesian adaptive design method for dose response curve is multiple endpoints oriented (Chang and Chow 2005). Thus, it can be used for various situations. The method can be improved by the following specific directions that (i) studying the relative merits and disadvantage of their method under various adaptive methods, (ii) examining the performance of an alternative method by forming the utility first with different weights to the response levels and then modeling the utility, and (iii) deriving sample size calculation formulas for various situations.

3.4.3 Adaptive Seamless Designs

As indicated earlier, an adaptive seamless design is a two-stage design that consists of two phases namely a learning (or exploratory) phase and a confirmatory phase. One of the major challenges for designs of this kind is that different study endpoints are often considered at different stages for achieving different study objectives. In

this case, the standard statistical methodology for assessment of treatment effect and for sample size calculation cannot be applied.

For a two-stage adaptive design, Bauer and Kohne (1994) proposed a method using Fisher's combination of independent p-values based on subsamples from different stages. Their method provides a great flexibility in the selection of statistical methods for hypothesis testing of subsamples. However, the choices for the stopping boundaries are not flexible to meet practical needs (Muller and Schafer 2001). As an alternative, it was proposed using linear combination of the independent p-values (Chang 2007). This method provides great flexibility in the selection of stopping boundaries, which can be calculated manually (Chow and Chang 2006). Chang's method, however, is valid under the assumption of constancy of the target patient populations, study objectives, and study endpoints at different stages. As pointed out earlier, it is most likely that the study objectives and study endpoints are different at different stages in practice. To have a valid and fair assessment of treatment effect based on combined data from the two stages, appropriate test statistics are necessarily developed.

Assuming that a two-stage adaptive seamless design utilizes two different study endpoints. At the first stage (learning or exploratory phase), the same study endpoint with a much shorter treatment duration or a biomarker is used. At the second stage (confirmatory phase), regular clinical study endpoint is used. Under the assumption that there exists a relationship between the two study endpoints, i.e., the first study endpoint is predictive of the second study endpoint, an appropriate test statistic can be developed. Let x_i be the observation of the study endpoint (e.g., biomarker) at the first stage from the ith subject, $i = 1, \ldots, n$ and y_j be the observation of the study endpoint (the primary clinical endpoint) from the jth subject, $j = 1, \ldots, m$. Assume that x_i's are independently and identically distributed with $E(x_i) = \nu$ and $Var(x_i) = \tau^2$; and $y_j's$ are independently and identically distributed with $E(y_j) = \mu$ and $Var(y_j) = \sigma^2$. Suppose that x and y can be related in a simple relationship as follows:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ε is an error term with zero mean and variance ξ^2. Furthermore, ε is independent of x. Thus μ can be estimated by $\hat{\mu} = w\hat{y} + (1 - w)\bar{y}$, where $\hat{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$, $\bar{y} = \frac{1}{m}\sum_{j=1}^{m} y_j$ and $0 \le w \le 1$. Note that $\hat{\mu}$ is the minimum variance unbiased estimator among all weighted mean estimators when the weight is given by $w = [n/(\beta_1^2\tau^2)]/[n/(\beta_1^2\tau^2) + m/\sigma^2]$. Based on the result of Khatri and Shah (1974), the variance of $\hat{\mu}$ can be approximated with a given bias order. Thus, under a given set of hypotheses (e.g., hypotheses for non-inferiority or equivalence) for evaluation of treatment effect can be derived accordingly.

In practice, when designing a clinical trial, the first question that the investigator will ask is how many subjects do we need in order to achieve the desired power at the pre-specified significance level. For a two-stage adaptive design with different study endpoints, sample size calculation also involves the allocation of sample sizes n and m at the two different stages. Let $m = \rho n$. Then the total sample size $N = (1 +$

$\rho)n$. For simplicity, consider testing the hypotheses for equality, n can be derived accordingly. Note that a two-stage adaptive seamless design can be viewed as a group sequential design with one planned interim analyses (Chow et al. 2017).

Remarks—In the previous section for illustration purpose, we only consider the case where the study endpoints are continuous variables. In practice, the study endpoints could be discrete (e.g., binary responses), time-to-event data, or mixed types of data. In these cases, similar idea can be carried out for development of appropriate statistical methods for analysis of the combined data under the assumption that there is a well-established relationship between the two study endpoints.
As indicated earlier, the traditional sample size calculation is often estimated by sample mean and sample variance from a small pilot study. Note that sample size calculation based on $s^2/\hat{\delta}^2$ is rather instable. As an alternative, it is suggested that the median of $s^2/\hat{\delta}^2$ be considered. As it can be seen the bias of the median approach can be substantially smaller than the mean approach for a small sample size and/or small effect size. However, in practice, we do not know the exact value of the median of $s^2/\hat{\delta}^2$. In this case, a bootstrap approach may be useful (see also Chow et al. 2017).

3.5 Obstacles of Retrospective Adaptations

In practice, retrospective adaptations such as adaptive-hypotheses are commonly encountered prior to database lock (or unblinding) and implemented through the development of statistical analysis plan. To illustrate the impact of retrospective adaptations, for simplicity, we will only consider the common situation for modifying hypotheses is switching a superiority hypothesis to a non-inferiority hypothesis. For a promising test drug, the sponsor would prefer an aggressive approach for planning a superiority study. The study is usually powered to compare the promising test drug with an active control agent. Let μ_T and μ_A be the mean efficacies of the test drug and the active control agent, respectively. Testing for superiority of the test drug over the active control agent amounts to testing the following hypotheses:

$$H_0 : \mu_T \leq \mu_A + \Delta \text{ versus } H_1 : \mu_T > \mu_A + \Delta \tag{3.7}$$

where $\Delta > 0$ is a known mixed superiority margin. If the null hypothesis H_0 in (3.7) is rejected, then we can conclude that the test drug is superior to the active control agent since μ_T is larger than μ_A by a fixed margin Δ.

However, the collected data may not support superiority. Instead of declaring the failure of the superiority trial, the sponsor may switch from testing superiority to testing the following non-inferiority hypotheses:

$$H_0 : \mu_T \leq \mu_A - \Delta \quad \text{versus } H_1 : \mu_T > \mu_A - \Delta. \tag{3.8}$$

If the null hypothesis H_0 in (3.8) is rejected, then we can conclude that the test drug is not worse than the active control agent by the margin Δ. Typically, the margin is carefully chosen so that $\mu_A - \Delta$ is larger than the placebo effect and, thus, declaring non-inferiority to the active control agent means that the test drug is superior to the placebo effect. The switch from a superiority hypothesis to a non-inferiority hypothesis will certainly increase the probability of success of the trial because the study objective has been modified to establishing non-inferiority rather than showing superiority. This type of switching hypotheses is accepted to the regulatory agency such as the U.S. FDA, provided that the impact of the switch on statistical issues (e.g., the determination of non-inferiority margin Δ) and inference (e.g., appropriate statistical methods) on the assessment of treatment effect is well justified.

To illustrate the concept of switching a superiority hypothesis to a non-inferiority hypothesis, we consider a parallel-group design with one interim analysis. Both groups for the test drug and active control agent have n patients at the interim analysis, and have additional n patients if the trial continues. For simplicity, assume that data are normally distributed with known variances σ_T^2 and σ_A^2 for the test drug and the active control agent, respectively. At the interim analysis, a superiority test rejects H_0 in (3.7) if $Z_1 > c_{1,\alpha}$, where $Z_1 = \sqrt{n}(\bar{d}_1 - \Delta)/\sqrt{\sigma_A^2 + \sigma_T^2}$, $\bar{d}_1$ is the difference between the sample mean of the test drug and the sample mean of the active control agent, α is a given significance level, and $c_{1,\alpha}$ is a constant specified according to (3.9). The trial continues if $Z_1 \le c_{1,\alpha}$. Let $Z_2 = \sqrt{2n}(\bar{d}_2 - \Delta)/\sqrt{\sigma_A^2 + \sigma_T^2}$, where $\bar{d}_2$ is the difference between the sample mean of the test drug and the sample mean of the active control agent, based on all patients at the end of the trial. The two-stage superiority test rejects H_0 in (3.7) if $Z_2 > c_{2,\alpha}$, where $c_{1,\alpha}$ and $c_{2,\alpha}$ are chosen so that

$$P\left(Z_1 > c_{1,\alpha}\right) + P\left(Z_2 > c_{2,\alpha}, Z_1 \le c_{1,\alpha}\right) = \alpha \quad \text{when H}_0 \text{ in (3.7) holds.} \tag{3.9}$$

Now, assume that the hypothesis of interest are switched to (3.8) when $Z_1 \le c_{1,\alpha}$ at the interim analysis. Let $\tilde{Z}_2 = \sqrt{2n}(\bar{d}_2 + \Delta)/\sqrt{\sigma_A^2 + \sigma_T^2}$. Consider the following test rule, which is referred to as a two-stage superiority-noninferiority test: If $Z_1 > b_{1,\alpha}$, conclude superiority ($\mu_T > \mu_A + \Delta$) and stop the trial; otherwise, continue the trial and if $\tilde{Z}_2 > b_{2,\alpha}$, conclude non-inferiority, but not superiority ($-\Delta < \mu T - \mu A \le \Delta$); otherwise, conclude not non-inferiority ($\mu T - \mu A \le -\Delta$). The constants $b_{1,\alpha}$ and $b_{2,\alpha}$ can be determined according to

$$P\left(Z_1 > b_{1,\alpha}\right) = \alpha \quad \text{when H}_0 \text{ in (3.7) holds} \tag{3.10}$$

$$P\left(Z_1 > b_{1,\alpha}\right) + P(\tilde{Z}_2 > b_{2,\alpha}, Z_1 \le b_{1,\alpha}) = \alpha \quad \text{when H}_0 \text{ in (3.8) holds}$$

It can be shown that $b_{1,\alpha}$ and $b_{2,\alpha}$ satisfying (3.10) exist; in fact, $b_{1,\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. Note that $c_{1,\alpha}$ satisfying (3.9) is larger than $b_{1,\alpha}$ satisfying (3.10). Note that the following pros and cons regarding the two

kinds of tests are observed. First, the two-stage superiority-noninferiority test has a better power at the interim stage than the two-stage superiority test. Second, the two-stage superiority-noninferiority test requires smaller sample size since the second stage test is for non-inferiority. Third, at the second stage, the two-stage superiority-noninferiority test can only conclude non-inferiority, whereas the two-stage superiority test may conclude superiority. Thus, in practice, it is worthy of pursuing the following research topics that (i) develop the two-stage superiority-noninferiority tests in more complex situations in terms of the study design and model assumption, (ii) consider a more flexible superiority-noninferiority test by adding a condition at the end of the first stage, i.e., if $Z_1 \leq b_{1,\alpha}$ but $Z_1 > a_{1,\alpha}$, then do not switch hypotheses; otherwise ($Z_1 \leq b_{1,\alpha}$), then switch hypotheses, (iii) derive formulas for sample size calculation. Since the hypotheses in two stages are different, it may not be a good idea to have the same sample size for each stage, and (iv) extend the results to group sequential tests with more than two stages.

Other Adaptive-Hypotheses—For clinical trials comparing several treatments or several doses of the same treatment with a placebo or an active control agent, a parallel-group design is usually considered. After the analysis of interim data, it is desirable to drop some treatment groups or dose groups showing no efficacy. It is also desirable to add some new treatment groups and/or modify the dose regimen for best clinical results. As a result, hypotheses have to be changed in the next stage of analysis.

To illustrate the concept of changing treatment arms, we consider the following simple case. Suppose that a group sequential design with $K = 2$ is adopted. At stage 1, two independent samples of size n are taken from a placebo control and a test drug with dosage x_1. Let $\bar{y}_{01}$ and $\bar{y}_{11}$ be the sample means from the control and test groups, respectively.

Assume $\bar{y}_{01}$ is distributed as $N(\mu_0, \sigma^2/n)$ and $\bar{y}_{11}$ is distributed as $N(\mu_1, \sigma^2/n)$. Let $Z_1 = \sqrt{n/2}(\bar{y}_{01} - \bar{y}_{11})/s_1$, where s_1^2 is the pooled sample variance. We reject the hypothesis H_{01}: $\mu_0 = \mu_1$ and we stop the trial if $|Z_1| > c_\alpha$. If $|Z_1| \leq c_\alpha$, then the trial continues with two independent samples of size n taken from the placebo control and the test drug with an increased dosage $x_2 > x_1$. At stage 2, the sample mean $\bar{y}_{02}$ from the control group is distributed as $N(\mu_0, \sigma^2/n)$, and the sample mean $\bar{y}_{22}$ from the test group is distributed as $N(\mu_2, \sigma^2/n)$. Hence, the null hypothesis at stage 2 is switched to H_{02}: $\mu_2 = \mu_0$. The test statistic at the second stage depends on what we assume of the relationship between the dosage and mean response. Without any assumption, we can consider the $Z_2 = \sqrt{3n/2}(\bar{y}_0 - \bar{y}_{22})/s_2$, where $\bar{y}_0 = (\bar{y}_{01} + \bar{y}_{02})/2$ and s_2^2 is the pooled sample variance based on 2 stages of data. We reject the null hypothesis H_{02} if $|Z_2| > b_\alpha$, where c_α and b_α are chosen so that

$$P(|Z_1| > c_\alpha) + P(|Z_2| > b_\alpha, |Z_1| \leq c_\alpha) = \alpha \tag{3.11}$$

when $\mu_0 = \mu_1 = \mu_2$ for a given significance level α. Note that μ_2 is estimated using the second stage data only, although μ_0 and σ^2 are estimated using data from both stages.

Assume now that the mean response and the dose level has the relationship that $\mu_k = \beta x_k$, $k = 1, 2$, where β is an unknown parameter. Then, μ_2 can be estimated using data from both stages. First, we estimate β by $\hat{\beta} = \frac{x_1 \bar{y}_{11} + x_2 \bar{y}_{22}}{x_1 + x_2}$. Then we estimate μ_2 by $\hat{\mu}_2 = \hat{\beta} x_2$. The test statistics Z_2 is modified as $Z_2 = (\bar{y}_0 - \hat{\mu}_2)/\sqrt{v}$, where v is an appropriate estimate of the variance of $\bar{y}_0 - \hat{\mu}_2$. We reject H_{02} if $|Z_2| > b_\alpha$, where c_α and b_α are still determined by (3.11) under $\mu_0 = \mu_1 = \mu_2$. Thus, it is worthy of pursuing the following research topics that (i) deriving valid statistical test procedures for more complex situations, such as designs with multiple dose levels and multiple interim analyses, (ii) employing Bayesian methods to relate dose levels with response means, and (iii) deriving sample size calculation formulas for various situations.

3.6 Concluding Remarks

As indicated earlier, although the use of adaptive design methods in clinical trials is motivated by its flexibility and efficiency, many researchers are not convinced and still challenge its validity and integrity (Tsiatis and Mehta 2003). As a result, many discussions are around the flexibility, efficiency, validity, and integrity. When implementing an adaptive design in a clinical trial, it is suggested a couple of principles that (i) adaptation should not alter trial conduct and (ii) type I error should be preserved must be followed when implementing the adaptive design methods in clinical trials (Li 2006). Following these principles, some basic considerations such as dose/dose regimen, study endpoints, treatment duration, and logistics should be carefully evaluated for feasibility (Quinlan et al. 2006). To maintain the validity and integrity of an adaptive design with complicated adaptations, it is strongly suggested that an independent data monitoring committee (IDMC) should be established. In practice, IDMC has been widely used in group sequential design with adaptations of stopping a trial early and sample size re-estimation. The role and responsibility of an IDMC for a clinical trial using adaptive design should clearly defined. IDMC usually convey very limited information to investigators or sponsors about treatment effects, procedural conventions, and statistical methods with recommendations in order to maintain the validity and integrity of the study.

When applying adaptive design methods in clinical trials, it is suggested that the feasibility of certain adaptations such as changes in study endpoints/hypotheses be carefully evaluated to prevent from any possible misuse and abuse of the adaptive design methods. For a complicated multiple adaptive design, it is strongly recommended that an independent data monitoring committee be established to ensure the integrity of the study. It should also be noted that although clinical trial simulation does provide a solution not the solution for a complicated multiple adaptive design. In practice, "how to validate the assumed predictive model for clinical trial simulation?" is a major challenge to both investigators and biostatisticians.

We are moving in the right direction and yet there is still a long way to go until we are able to address all of the scientific issues from clinical, statistical, and regulatory perspectives as described earlier. Detailed design-specific guidances (e.g., guidances regarding sample size calculation/allocation and statistical/clinical considerations for a two-stage phase I/II or phase II/III seamless adaptive trial design) must be developed by the regulatory agencies before implementation of adaptive design methods in pharmaceutical/clinical research and development. In addition, qualification, composition, role/responsibility, and function/activity of an independent data monitoring committee for implantation of adaptive trial design need to be established for an objective and unbiased assessment of the treatment effect of the drug under investigation. Thus, from future perspectives, it is suggested that the escalating momentum for the use of adaptive design methods in clinical trials proceed with caution. At the same time, valid statistical methods for interested adaptive designs with various adaptations should be developed to prevent the possible misuse and/or abuse of the adaptive design methods in clinical trials. More details regarding recent development of statistical methodologies for specific adaptive designs such as adaptive dose finding, genomic-guide target clinical trial design, and two-stage adaptive seamless (phase I/II or phase II/III) designs can be found in Pong and Chow (2010) and Chow and Chang (2012).

References

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs*. New York: Oxford University Press.

Babb, J. S., & Rogatko, A. (2001). Patient specific dosing in a cancer phase I clinical trial. *Statistics in Medicine, 20,* 2079–2090.

Bauer, P., & Kieser, M. (1999). Combining different phases in development of medical treatments within a single trial. *Statistics in Medicine, 18,* 1833–1848.

Bauer, P., & Kohne, K. (1994). Evaluation of experiments with adaptive interim analysis. *Biometrics, 50,* 1029–1041.

Bauer, P., & Rohmel, J. (1995). An adaptive method for establishing a dose-response relationship. *Statistics in Medicine, 14,* 1595–1607.

Brannath, W., Koening, F., & Bauer, P. (2003). Improved repeated confidence bounds in trials with a maximal goal. *Biometrical Journal, 45,* 311–324.

Branson, M., & Whitehead, W. (2002). Estimating a treatment effect in survival studies in which patients switch treatment. *Statistics in Medicine, 21,* 2449–2463.

Chang, M. (2007a). Adaptive design method based on sum of p-values. *Statistics in Medicine, 26,* 2772–2784.

Chang, M. (2007b). *Adaptive design theory and implementation using SAS and R*. Taylor and Francis, New York: Chapman and Hall/CRC Press.

Chang, M., & Chow, S. C. (2005). A hybrid Bayesian adaptive design for dose response trials. *Journal of Biopharmaceutical Statistics, 15,* 667–691.

Charkravarty, A. (2005). Regulatory aspects in using surrogate markers in clinical trials. In T. Burzykowski, G. Molenberghs, & M. Buyse (Eds.), *The evaluation of surrogate endpoint*. Springer.

Chow, S. C., & Chang, M. (2006). *Adaptive design methods in clinical trials*. New York: Chapman and Hall/CRC Press; Taylor and Francis.

Chow, S. C., & Chang, M. (2012). *Adaptive design methods in clinical trials* (2nd ed.). New York: Chapman and Hall/CRC Press; Taylor and Francis.
Chow, S. C., Chang, M., & Pong, A. (2005). Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics, 15,* 575–591.
Chow, S. C., & Liu, J. P. (2003). *Design and Analysis of Clinical Trials* (2nd ed.). New York: Wiley.
Chow, S. C., Lu, Q., & Tse, S. K. (2007). Statistical analysis for two-stage adaptive design with different study points. *Journal of Biopharmaceutical Statistics, 17,* 1163–1176.
Chow, S. C., & Shao, J. (2005). Inference for clinical trials with some protocol amendments. *Journal of Biopharmaceutical Statistics, 15,* 659–666.
Chow, S. C., & 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., Wang, H., & Lokhnygina, Y. (2017). *Sample size calculation in clinical research* (3rd ed.). New York: Chapman and Hall/CRC Press; Taylor and Francis.
Christensen, R. (1996). Exact tests for variance components. *Biometrics, 52,* 309–314.
Chung-Stein, C., Anderson, K., Gallo, P., & Collins, S. (2006). Sample size reestimation: A review and recommendations. *Drug Information Journal, 40,* 475–484.
Crowley, J. (2001). *Handbook of statistics in clinical oncology*. New York: Marcel Dekker Inc.
Cui, L., Hung, H. M. J., & Wang, S. J. (1999). Modification of sample size in group sequential trials. *Biometrics, 55,* 853–857.
Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika, 58,* 403–417.
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. (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.
Feng, H., Shao, J., & Chow, S. C. (2007). Group sequential test for clinical trials with moving patient population. *Journal of Biopharmaceutical Statistics, 17,* 1227–1238.
Friede, T., & Kieser, M. (2004). Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics, 3,* 269–279.
Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., & Pinheiro, J. (2006). Adaptive design in clinical drug development—an executive summary of the PhRMA working group (with discussions). *Journal of Biopharmaceutical Statistics, 16*(3), 275–283.
Gallo, J., & Khuri, A. I. (1990). Exact tests for the random and fixed effects in an unbalanced mixed two-way cross-classification model. *Biometrics, 46,* 1087–1095.
Gasprini, M., & Eisele, J. (2000). A curve-free method for phase I clinical trials. *Biometrics, 56,* 609–615.
Hardwick, J. P., & Stout, Q. F. (2002). Optimal few-stage designs. *Journal of Statistical Planning and Inference, 104,* 121–145.
Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal, 43,* 581–589.
Hommel, G., Lindig, V., & Faldum, A. (2005). Two stage adaptive designs with correlated test statistics. *Journal of Biopharmaceutical Statistics, 15,* 613–623.
Hung, H. M. J., Cui, L., Wang, S. J., & Lawrence, J. (2005). Adaptive statistical analysis following sample size modification based on interim review of effect size. *Journal of Biopharmaceutical Statistics, 15,* 693–706.
International Conference on Harmonization Guideline E10: Guidance on Choice of Control Group and Related Design and Conduct Issues in Clinical Trials. The United States Food and Drug Administration, Rockville, Maryland, July 2000.
Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. New York, NY: Chapman and Hall.
Jennison, C., & Turnbull, B. W. (2005). Meta-analysis and adaptive group sequential design in the clinical development process. *Journal of Biopharmaceutical Statistics, 15,* 537–558.

Kelly, P. J., Sooriyarachchi, M. R., Stallard, N., & Todd, S. (2005a). A practical comparison of group-sequential and adaptive designs. *Journal of Biopharmaceutical Statistics, 15,* 719–738.

Kelly, P. J., Stallard, N., & Todd, S. (2005b). 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.

Khatri, C. G., & Shah, K. R. (1974). Estimation of location of parameters from two linear models under normality. *Communications in Statistics, 3,* 647–663.

Khuri, A. I., Mathew, T., & Sinha, B. K. (1998). *Statistical tests for mixed linear models*. New York, NY: Wiley.

Lachin, J. M. (1988). Statistical properties of randomization in clinical trials. *Controlled Clinical Trials, 9,* 289–311.

Lan, K. K. G., & DeMets, D. L. (1987). Group sequential procedures: Calendar versus information time. *Statistics in Medicine, 8,* 1191–1198.

Lee, Y., Wang, H., & Chow, S. C. *A bootstrap-median approach for stable sample size determination based on information from a small pilot study*. Unpublished manuscript.

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

Li, N. (2006). Adaptive trial design—FDA statistical reviewer's view. Presented at the CRT2006 Workshop with the FDA, Arlington, Virginia, April 4, 2006.

Li, G., Shih, W. J., & Wang, Y. (2005). Two-stage adaptive design for clinical trials with survival data. *Journal of Biopharmaceutical Statistics, 15,* 707–718.

Liu, Q., & Chi, G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics, 57,* 172–177.

Liu, Q., Proschan, M. A., & Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. *Journal of American Statistical Association, 97,* 1034–1041.

Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., & Krams, M. (2006). Adaptive seamless phase II/III designs—background, operational aspects, and examples. *Drug Information Journal, 40,* 463–474.

Mugno, R., Zhus, W., & Rosenberger, W. F. (2004). Adaptive urn designs for estimating several percentiles of a dose response curve. *Statistics in Medicine, 23,* 2137–2150.

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

Ofversten, J. (1993). Exact tests for variance components in unbalanced mixed linear models. *Biometrics, 49,* 45–57.

O'Quigley, J., Pepe, M., & Fisher, L. (1990). Continual reassessment method: A practical design for phase I clinical trial in cancer. *Biometrics, 46,* 33–48.

O'Quigley, J., & Shen, L. (1996). Continual reassessment method: A likelihood approach. *Biometrics, 52,* 673–684.

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

Pong, A., & Chow, S. C. (2010). *Handbook of adaptive designs in pharmaceutical and clinical development*. New York: Taylor & Francis.

Posch, M., & Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal, 41,* 689–696.

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

Posch, M., Konig, F., Brannath, W., Dunger-Baldauf, C., & Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine, 24,* 3697–3714.

Proschan, M. A. (2005). Two-stage sample size re-estimation based on a nuisance parameter: A review. *Journal of Biopharmaceutical Statistics, 15,* 539–574.

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

Proschan, M. A., Leifer, E., & Liu, Q. (2005). Adaptive regression. *Journal of Biopharmaceutical Statistics, 15,* 593–603.
Quinlan, J. A., Gallo, P., & Krams, M. (2006). Implementing adaptive designs: logistical and operational consideration. *Drug Information Journal, 40,* 437–444.
Rosenberger, W. F., & Lachin, J. (2002). *Randomization in clinical trials*. New York: Wiley.
Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., & Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics, 57,* 909–913.
Sampson, A. R., & Sill, M. W. (2005). Drop-the-loser design: Normal case (with discussions). *Biometrical Journal, 47,* 257–281.
Shao, J., Chang, M., & Chow, S. C. (2005). Statistical inference for cancer trials with treatment switching. *Statistics in Medicine, 24,* 1783–1790.
Shen, Y., & 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.
Tsiatis, A. A., & Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika, 90,* 367–378.
Wang, S. J., & Hung, H. M. J. (2005). Adaptive covariate adjustment in clinical trials. *Journal of Biopharmaceutical Statistics, 15,* 605–611.
Wang, S. J., O'Neill, R. T., & Hung, H. M. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics, 6,* 227–244.
Wang, S. K., & Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for a sequential trials. *Biometrics, 43,* 193–200.
Wei, L. J. (1978). The adaptive biased-coin design for sequential experiments. *Annal of Statistics, 9,* 92–100.
Whitehead, J. (1997). Bayesian decision procedures with application to dose-finding studies. *International Journal of Pharmaceutical Medicine, 11,* 201–208.
Woodcock, J. (2005). FDA introduction comments: Clinical studies design and evaluation issues. *Clinical Trials, 2,* 273–275.
Zhang, W., Sargent, D. J., & Mandrekar, S. (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine, 25,* 2365–2383.

Chapter 4
Best Practices in Clinical Trial Simulations for Adaptive Study Designs

Cristiana Mayer and J. Kyle Wathen

4.1 Introduction

If asked to explain what modeling and simulation activities are about and where they are used, simulation of virtual clinical trials in the pharmaceutical industry is not the first example that comes to mind. It seems more natural to imagine aerospace engineers who model the aerodynamic properties of a new aircraft and simulate its performance under different flying conditions. One may also think of the National Aeronautics and Space Administration which has been using simulations for decades for astronauts training. Entertainment will likely be another common example where simulating a virtual reality has played a big role in the production of movies and games. Why not the pharmaceutical industry then?

Pharmaceutical research and development (R&D) is a scientific field notoriously known to be challenged by high failure rates and continuously increasing costs in the attempt of bringing better and safer medicines to the market. Ignoring large differences by therapeutic class, the overall clinical approval success rate between 1999 and 2004 was estimated to be 19% based on recent research data on the investigational compounds of the 50 largest pharmaceutical companies, whose size was determined by their sales in 2006 (DiMasi et al. 2010).

It would be obvious that in such high-risk environment the business model for the pharmaceutical industry would have routinely included a heavy-duty use of modeling and simulations before investing in multi-million dollar clinical studies, especially in the early learning phases of development. But this did not materialize before the world of adaptive and other innovative clinical trials stepped in the pharmaceutical

C. Mayer (✉) · J. K. Wathen (✉)
Quantitative Sciences, Janssen Research & Development LLC, 1125 Trenton-Harbourton Road, 08560 Titusville, NJ, USA
e-mail: CMayer1@its.jnj.com

J. K. Wathen
e-mail: kwathen@its.jnj.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_4

R&D environment. This chapter is intended to illustrate the simulation process in the context of pharmaceutical drug development with emphasis to the design of adaptive clinical trials. The definition of adaptive trials and the most common types of adaptive designs are left to other publications (Bauer et al. 2016; He et al. 2014).

In the next section, the motivation for the new trends in planning, conducting and sharing modeling and simulation activities is described. Section 4.3 illustrates the simulation process and describe in details each of the different components that constitute a simulation study. Section 4.4 describes the scenario planning followed by testing in Sect. 4.5. The chapter ends with the benefits and challenges in Sect. 4.6. and the overall conclusions in Sect. 4.7.

4.2 Motivation

The idea of shifting the R&D paradigm and modernizing drug development is not new. More than a decade ago both industry and Health Authorities recognized that drug development was stagnating and had to transform to thrive. This necessity was described already in the FDA report (2004) on "Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products" and measures to inject innovation presented in the subsequent "Innovation/Stagnation: Critical Path Opportunity List" (2006) as well as in the EMEA Innovation Task Force and concept papers (2007, Innovative Drug Development Approaches; 2011, EMA Report on the implementation; 2010 EMA Road Map to 2015).

The introduction of novel statistical methodologies for adaptive trial designs and their broader utilization have vehemently flourished in the last decade. Clinical trials are designed and executed to gather reliable scientific evidence that allows for quantitative-based sound decision making as early as possible, because "failing early" is a success by saving time and resources (Kannt and Wieland 2016). Much has been written on how to improve the R&D business model and it is nowadays common to prefer the drug development paradigm labeled "quick win, fast fail" as illustrated in Paul et al. (2010). According to this paradigm, a smaller number of compounds advance in Phase 2b, and Phase 3 but those that advance have a higher probability of successfully being approved and launched on global markets.

Nowadays the current business model in the pharmaceutical industry must be strongly anchored on quantitative approaches for evaluating and comparing the designs and analysis methods of clinical trials to enable better decision-making and quantify more precisely risks as well as predict more precisely the outcomes. The need to make the "right" decision at the "earliest" time point requires simulations that allow a detailed evaluation of outcome uncertainty per unit of time and of costs.

Given the richness of recent medical, statistical and computing technological advances, adaptations in clinical trials triggered by accumulating data in near real time has brought to the spotlight the importance and usefulness of modeling and simulation. Not only adaptive design of clinical trials, but also the combination of different statistical approaches, like Bayesian and frequentist concepts, and advanced

modeling techniques, like disease progression models or model-based meta-analyses, are more commonly and routinely used. All these approaches require careful and detailed statistical evaluations which cannot avoid simulations.

The interest and support from Health Authorities has also boosted the interest in and application of clinical trials simulations (Westfall et al. 2008). The regulators acknowledge the importance of clinical trial simulation in the armamentarium of tools for innovation in R&D to help predict efficient designs for development programs that reduce the number of trials and patients, improve decisions on dosing, and increase informativeness (2006 Challenge and Opportunity on the Critical Path to New Medical Products). The efficient and intelligent use of the simulation tools on an adaptive design clinical trial should ultimately help increase the likelihood of success at the program development level (Wang 2009). As stated in the FDA guidance "Computer simulations can play a crucial role in adaptive designs and can provide the operating characteristics of the study design under different scenarios" (FDA Guidance on Adaptive Designs for Medical Device Clinical Studies p. 28, July 2016).

In December 2016, the US Congress signed into law "The 21st Century Cures Act". Under title III—Subtitle C "Modern trial design and evidence development", direct reference to the use of more complex adaptive and other novel trial designs is made paired with the need for a dialogue between the FDA and sponsors on the expectations and technical issues related to modeling and simulations.

The Prescription Drug User Fee Act (PDUFA) V sunsets on September 30, 2017. In the current PDUFA VI legislation approved by the United States Congress in August 2017, an enhancement is focused on promoting simulation approaches considered a critical tool to support innovation and regulatory statistical sciences. The legislation mandates the necessity to clarify for sponsors the regulatory expectations around simulations studies that aim to adequately characterize the performance of more complex trials.

Beyond the regulatory and legislative changes in the last few years, the expansion of statistical methodologies on adaptive clinical trial designs and advancement in computing technologies and software development to support the design and execution of more sophisticated and complex adaptive study designs have accelerated the use of simulations in the context of drug development.

As Ruberg (2016) has eloquently hypothesized, "The prediction that the future will use more adaptive clinical trials may seem to be an easy one given the burgeoning efforts that are underway today. […] This will be a fundamental change from operating in "batch mode," whereby clinical trials proceed in a step-by-step fashion with discrete decision points—accompanied by "white space" in between completion of trials and initiation of the next trial—to operating in "interactive mode" in which predefined decision criteria supported by models and simulations allow valid statistical inference" (page 60). Adaptive designs are indeed an invaluable tool to increase the efficiency of drug development process by utilizing accumulating information to make decisions faster on a compound's safety or efficacy (Gallo et al. 2006).

Another important reason that significantly promote the use of simulation techniques in the pharmaceutical industry is to bridge gaps in knowledge for clinical

testing of investigational drugs while quantifying the involved risks. If, on one side, an adaptive design may be chosen because improves the information quality and value for the money invested in a resource-constrained environment (maximize the high-quality information amount per dollar spent), simulations are the critical tool to gain a better understanding of what can be expected in a trial, what type of efficiencies and potential reduction in failure rate can be achieved and the degree of uncertainty around such quantities. In addition, simulations satisfy the need to explore a large set of scenarios to pressure-test the clinical study design and to explain key operating characteristics (OC) and statistical properties that often cannot be derived analytically.

4.3 Simulation Process

In the attempt to define the term "simulation" in the context of clinical trials, one may describe it as the utilization of computer intensive procedures involving mathematical and statistical techniques for conducting virtual clinical experiments on the computer. The assessment of the performance of a variety of statistical and mathematical methods, clinical trial designs, and/or modeling approaches cannot be achieved by conducting clinical studies in the real world. Consequently, a computer simulation is an attempt to model a real-life clinical study or a hypothetical situation within a pharmaceutical compound development program on a computer so that it can be studied to see how the trial and/or statistical or mathematical model would work under a wide range of plausible—not all necessarily likely—scenarios. By changing values of key variables, assumptions and scenarios, predictions can be made about the performance of the study design or compound development program features.

More concisely, clinical trials are conducted to test the safety and efficacy of new medicinal products and devices whereas a clinical trial simulation is conducted to understand how the design will perform in practice and why the final design was selected.

4.3.1 Simulation Terminology

For ease of explanation, the simulation terminology that is used throughout this chapter is introduced first. Without loss of generality, concepts are explained in the context of a two-arm trial comparing the standard of care (S) to an experimental (E) treatment. The approaches of simulation discussed here can easily be adapted to fit more complex settings.

There are numerous commercial and freeware packages available for clinical trial simulation. Among the more frequently used commercial packages are: (1) EAST® 6, (2) ADDPLAN®, (3) COMPASS® 2.0 and (4) FACTS. The M. D. Anderson software download site (https://biostatistics.mdanderson.org/SoftwareDownload/) provides a

wide variety of free software for simulating Bayesian designs. Quite often custom simulation software is developed to make the design fit the trial goals rather than changing the trial to match available designs in the commercial software. It is not uncommon that custom simulation software is created to address the adaptive elements, logistical considerations or other decision points that are being considered. Regardless of what software is used, the term **Virtual Trial Simulator** (VTS) is used to reference the complete collection of components necessary to simulate and understand the clinical trial design. The VTS consists of a virtual trial, analysis model(s), decision rules, the simulation model, scenarios and performance metrics.

The clinical trial being simulated is called the virtual trial. The **virtual trial** defines the design and consists of the analysis model and decision rules. The goal of the simulation is to utilize software to make the virtual trial design match exactly, or as close as possible, what will be done when the clinical trial is conducted. Using this approach, the virtual trial is simulated repeatedly by enrolling virtual patients, or computer generated patients, for the study team to obtain a better understanding of how the design will perform in the real world. The OCs are used to characterize the design and consist of frequently required statistical quantities such as false-positive error rate, power and average sample size. Before the virtual trial can be simulated the team must specify collection of scenarios. A **scenario** is a list of values specifying the "true" underlying parameters, such as true response rates for S and E or the true recruitment rate. The **simulation model** is the component that describes how the virtual patient data, such as outcome(s), enrollment time and patient characteristics are simulated.

There are two methods for simulating a clinical trial: (1) Commercial software or (2) custom software. There are advantages and disadvantages to each approach and the decision to use one over the other can be rather easily determined by answering two questions: (1) Does commercial software exist that very nearly fits the design for trial or would major simplification need to be made to use the available packages and (2) Does the team have sufficient time and the required skills/ability to develop custom software that can better match the trial design while meeting the timelines often dictated by completely other reasons in drug development.

It becomes apparent that the complexity and desirability of the study design must be evaluated in conjunction with the skills and abilities of staff to create specific simulation software. In the end, it is the tradeoff between the desire for the simulation to match the real clinical trial versus using a pre-made design and the resources available to the project team that will dictate what choice is to be made among the VTS tools.

Commercial software will often provide a wide variety of additional information that is used in understanding the performance of different trial designs. The obvious appeal for using the commercial software is the convenience of having pre-prepared tools at disposal to facilitate the comparison of multiple trial designs which are reproducible and require a relatively short development time. In addition, multiple users can access and utilize the same tools without the needs for experience and knowledge of computer science programming languages or computing technologies beyond the software specifications or input information associated with the specific

design engine of choice. The biggest drawback with commercial software is the lack of extensibility to create new designs that are not currently in the software, regardless of how close the current options match the desired design features.

A major advantage of **custom software** is the ease of obtaining any summary information that could be used for decision making, such as values that are not routinely examined. In addition, with custom simulation software one can often go well beyond the OCs and average behavior to better understand what could potentially happen when the real trial is executed. The biggest advantage of custom software is the complete flexibility to accommodate any design so that the design can match the trial goals and features. The process of developing custom simulation software often leads to a much deeper and extensive understanding of what could happen during the clinical trial execution. The biggest disadvantage to custom simulation software is that it requires developers with expertise and knowledge of clinical trial designs and of programming code, time to create the code with the necessary flexibility to be used in different settings and time to validate the code to a reasonable degree of comfort.

The goal of a simulation is to determine the best design options for conducting the trial, especially in the context of adaptive designs where adaptations are planned to potentially offer different outcomes along the way, such as an early futility decision, a sample size readjustment or treatment dose or regimens selection in multiple stages. The simulation may also try several different types of analysis model and decision criteria. In addition, the simulation may include other aspects such as multiplicity adjustment strategies or missing data handling methods. In many settings addressing all these questions analytically may be difficult or even impossible. In the following section the various components of the VTS are described as well as an overview of the simulation process.

4.3.2 Analysis Models and Assumptions

The analysis model is the component of the VTS that most project teams in the pharmaceutical industry are familiar with. The analysis model is typically described in the protocol and details given in the Statistical Analysis Plan (SAP). The analysis model may often make assumptions, like normality of the data or proportional hazard function, that may not be true.

There are many features of the analysis model that the simulation will be designed to test. For example, if historical data is available to suggest that the hazard for patients with a disease is decreasing over time and that based on the historical treatments the proportional hazard assumption is unlikely to hold, then a simulation could test how robust the proposed analysis model is to departures from the underlying assumptions. In this setting, there are various approaches to how to analyze the data and ultimately make decisions. In practice one would provide all the details of analysis options. However, without loss of generality, only two options for analyzing the data are presented here and labeled as Option 1 and Option 2. After the project team

has identified a simulation model that would simulate virtual patients that closely resemble real patients the VTS can be used to compare Option 1 and Option 2. This would allow the team to compare the likelihood of success (or other OCs) for the two different options. It is a good practice to assume that the simulation model and analysis model are not the same and simulate the trial under various simulation models to understand the risk, and generate a robust set of scenarios, as addressed in the Scenario Planning Sect. 4.4. One of the goals of the simulation is to make sure that the analysis model will provide the best chance of success under a variety of simulation models, or at least many scenarios.

In combination with the analysis model the decision rules define what actions will be taken based on the results of the analysis. For example, in a frequentist analysis a p-value <0.05 may lead one to conclude "success" or in a Bayesian design a posterior probability that the experimental treatment is better than control of a very small amount may lead to a futility conclusion. It is important to make sure that the decision rules are in line with clinical intuition and practical considerations. Statistical quantities are to be molded to conform and satisfy clinical meaningful decisions and contribute to a scientific based quantitative decision making process that also combines elements of commercial, regulatory and clinical value.

4.3.3 Simulation Model

A subtle, but important, aspect of simulating a clinical trial is to understand the difference between the analysis model and the simulation model. Statisticians and clinical teams are familiar with the analysis model, because the analysis model is used during interim analysis and final analysis to analyze the data obtained from the clinical study and are typically documented in the trial SAP. Conversely, the simulation model is not part of a conventional fixed design because the OCs, such as power and false-positive error rate, can be obtained theoretically with no need for a simulation study. That is, in a standard design, one often relies on large sample theory and assumes the simulation model and analysis model are the same, thus no need to specify any simulation model.

In the context of adaptive studies and other more sophisticated trial designs, to simulate the trial, the simulation model provides the details about how the trial will be simulated. This may sound trivial but it is experience of the authors that the simulation process may be somewhat overlooked. The details include specifics about the assumptions made to generate virtual patients and the logistical aspects of the trial used in the simulation. For virtual patients, the simulation model includes details about how the patient outcomes and patient characteristics, if needed, will be generated during the simulation. In addition, the simulation model will specify the treatments impact on several patient outcomes. It is important to note that how the patient outcomes are simulated does not have to match the analysis model. For example, the simulation model may include a patient covariate, such as age, that impacts the primary patient efficacy variable but the patient's age in not part of the

analysis model. If the analysis model and the simulation model do not match, then the simulation will highlight how sensitive the analysis model and decision making rules are to departures from what is expected or even planned. It is a best practice to consider cases where the simulation model and analysis models do not match.

One of the key components of the simulation model is the enrollment pattern and patient arrival times in the virtual trial. In practice, it is common for the patient accrual to increase over the first several months of the trial and often accelerate even faster towards the end, especially when competing enrollment is used among sites. The details on how this would be translated in simulation constitutes part of the simulation model. While it may take more effort to get the virtual trial to match the expectations of the real trial, simulations can often identify potential problems before they occur. In the context of different simulated accrual rates, for example, the ramp-up in accrual can show if an interim analysis may be planned when too little data will be available or it is not feasible operationally if patients' accrual will be fast relative to the timepoint of the outcome assessment. A commonly used approach is to assume patients arrive according to a Poisson process where the rate is either fixed or changes over time, see Fig. 4.1 as an example. This component of the VTS can be used to graphically illustrate potential different accrual patterns with the associated confidence bands on top of the timing when data would likely be available. Figure 4.1 provides an example of patient accrual where both the number of patients recruited per site and the number of sites in the study increases at the beginning of the study. In addition, Fig. 4.1 shows how much patient data would be available over the time horizon, assuming that the outcomes of interest are observed at 6 and 12 week time point after treatment. This type of plot can be very helpful for planning purposes and can easily convey when the data would be available for decision making at various time points in the trial.

4.3.4 Performance Metrics

Performance metrics are the key deliverable of a simulation study. They are used to explain how a design will perform in practice and are critical for comparing design options. The amount of information that is captured during the simulation can help to provide a clear understanding of both the statistical properties, as well as highlight logistical and cost/time characteristics.

There are many situations where the usual OCs like power, false-positive error rate and average sample size are not sufficient to provide a clear picture for comparing among various designs. New methodology in the field of adaptive clinical trial designs is developed with a specific goal in mind. For example, outcome adaptive randomization is an approach often used to unbalanced randomization to favor the treatment that, on average, has better performance based on the accruing patient information. In this setting it is typical to report the usual OCs mentioned above. However, it is also important to report additional metrics such as the probability that

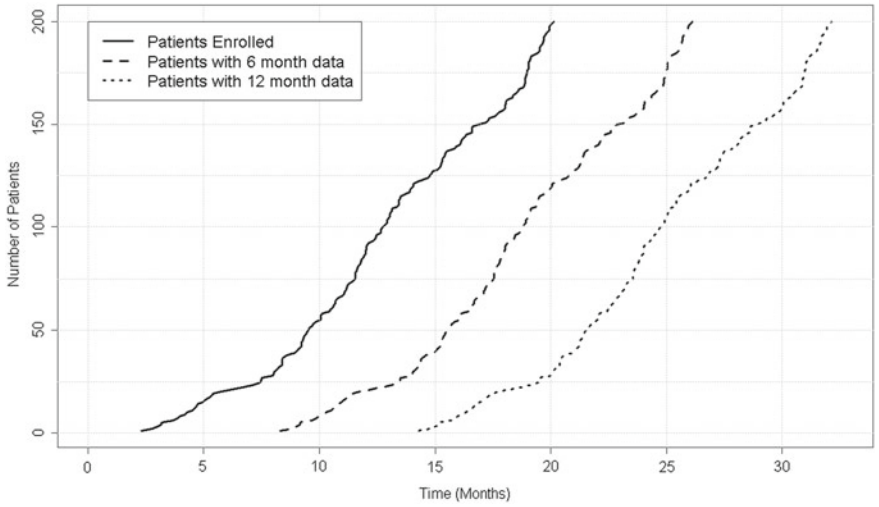


Fig. 4.1 Example trial accrual and patient data availability

more patients are assigned to the inferior treatment and the sample size distribution per arm.

As a second example, consider a trial that is intended to gain an understanding of a dose-toxicity curve as well as a dose-response curve and the tradeoff between them. In this example, a simulation would want to report the likelihood of selecting each of the doses under a variety of scenarios, (see the next section for more about scenario planning in this example). In addition, a simulation should report quantities to summarize the likelihood of selecting safe and/or effective doses as defined by the trial objectives and the average, minimum and maximum number of patients treated on each dose. It is a very good practice to capture several examples of individual virtual trials to show to the team how different decisions could be made under certain circumstances.

A third example of performance metrics that can be very useful are non-statistical quantities. For example, it is often useful to compare design strategies based on metrics such as financial considerations, time to market (even if an adaptive trial may take longer to complete, is the drug development program significantly faster overall compared to others?), or selection of subpopulations that may benefit for the treatment that maximizes the benefit-risk ratio.

Making general statement about the OCs beyond the usual false-positive error rate and statistical power is difficult because the desired OCs are often highly dependent on the specifics of the design and settings that the simulation is being performed under. It is critical to perform an extensive simulation study when trying a new statistical methodology and utilize virtual patients, rather than real one, to determine whether the new statistical methods provide the desired performance.

4.4 Scenario Planning

Scenario planning can be thought of as an opportunity to examine "what if" cases. It is common for team members or others, such as senior management, to inquire about how a design performs in specific settings. For example, the following "what if" questions could be raised: (1) what if the recruitment rate is much slower/faster than expected?, (2) what if the dropout rate is higher than expected?, (3) what if the control/placebo effect is different than was observed in historical trials?, (4) what if only patients under age 50 respond to our treatment?, or (5) What if multiple interim analyses are considered instead of just one?

As a team develops a trial and identifies a set of "what if" questions, scenarios can be created to answer each of the questions. Being able to answer such questions is a major benefit of the simulations. Typically, one simple question like those listed above can lead to several subsequent questions that ultimately result in a separate collection of scenarios. Each scenario describes a potential "true state of nature" and thus simulating many scenarios allows study teams to understand how a design could ultimately perform. In scenario planning the team thinks carefully about the question, determines which parameters need to be varied to answer the question, then simulates the scenarios to address the question.

To help clarify what scenario planning entails, a few of the most common "what if" questions listed above are used to describe what a team may consider when they are planning scenarios. Each example provides a few thoughts the team may go through and how scenarios could be built, to answer the questions. Typically, the main concern is how much, if any, the changes will impact the OCs or likelihood of success.

Example 1
Question: What if patient recruitment is much slower than expected?
Scenario Planning: The team has provided the baseline assumptions around patient accrual and simulations have been conducted assuming the recruitment follows a Poisson process with the anticipated recruitment rate. It is a best practice to consider a recruitment rate that may vary from 25 to 200% of what is anticipated to help the team understand the impact on the design. It is often informative to consider alternatives where the recruitment rate ramps up at the beginning of the study.

Example 2
Question: What if placebo (or control) response rate is different than historical?
Scenario Planning: The team has provided the baseline assumptions around how patients will respond to placebo (or control). These estimates are often based on historical data, literature or expert opinion. It is a best practice to consider varying the placebo rate and determine the impact on the design. In Bayesian designs this can be a very important concern, especially if an informative prior is utilized. Typically, the guess on the range on the parameter of interest has scientific validity, but it still a good idea to simulate some scenarios where the true parameter of interest is outside the provided range.

Example 3
Question: What if an additional interim analysis (IA) is included?
Scenario Planning: This questions arises often in the context of adaptive trials as the perception is that looking more frequently can lead to better performance. Simulating various cases where more IAs are included can provide the team with a clear understanding of the benefits of more frequent monitoring and trade-off with some measure of penalty. If the initial design has one IA and a final analysis (FA) then it would be very helpful to consider the potential gain of several options: (1) adding one additional IA midways between the original IA and FA, (2) consider moving the IA earlier/later or (3) frequent/continuous monitoring such as monthly monitoring. It would be important to report the impact on the OCs such as increase in false-positive error rate or statistical power as well as the changes in other quantities of interest such as probability of stopping early for futility under the null hypothesis or success under an alternative hypothesis. One key factor to keep in mind in addition to the usual Type 1 error rate control is what the practical limitations are around doing more frequent monitoring from an operational perspective.

4.5 Testing

For both commercial and custom software, a fair amount of testing should be performed. However, the testing that is done is highly dependent on the software category (commercial or custom). For commercial software, it important to make sure to "test" that the software is applied to the correct setting matching the design and features the researcher has chosen to test. In addition, it is important to make sure to correctly input the information (e.g. parameter defining the software specifications) to match the assumptions made by the study team and have a clear understanding of how such parameter/feature entries impact the design. It is important to ensure that the simulated data, models and results are accurate and match the desired assumptions to meet the objectives of the simulation and answer the questions that the simulation was designed to address. It is also helpful to simulate some scenarios at extremes to check how the results match with what is expected. For custom code, testing should be more extensive and include the steps suggested for the commercial software packages. However, explaining the ideal testing environment for custom code is well beyond the scope of this chapter, as there are no prespecified fields to be populated for a given design engine. Lastly, the number of simulations must be carefully planned, as the variance of quantities of interest resulting from the simulation study greatly changes with the number of simulation runs per scenario. One cannot sacrifice accuracy of results for practical execution and computing time.

4.6 Benefits and Challenges

In the conventional fixed design study, it is often straightforward to obtain the information needed to determine the adequate sample size for a given expected treatment effect. There are many tools available to answer the simple, and common, question "What sample size do I need to achieve this primary objective with sufficient statistical power?".

When looking at the landscape of adaptive clinical trials with multiple decision rules and adaptations, multiple looks at the data and multiple objectives, the assumptions made to conduct the simulation are an integral part of the dialogue between functions in clinical development. It is no longer a matter of a "simple" statistical sample size calculation. It is now a matter of combining assumptions around the clinical and commercial value of different study designs, the clinical operation aspects in conducting a certain trial design that have an impact on the scenario analysis as well the statistical methodologies and quantities to be compared among studies.

Related to the collection of assumptions, statisticians who perform simulations may encounter the challenge of engaging colleagues in other functions who are not so familiar with the concepts around adaptive designs and the simulation process. It is often conceivable that some of the clinical team members may not completely understand what simulations entail. The interaction among statisticians and other team members is an aspect of the conduct of the simulation process that cannot be emphasized enough. The dialogue between statistician and study team consists of 2 parts: eliciting information for constructing meaningful and plausible assumptions, and providing a clear description of the design and scenarios of interest to assess how it matches the clinically meaningful objectives. The benefit of obtaining a better and more comprehensive understanding of the trial features and performance is not always an obvious benefit to non-statisticians. Colleagues in the clinical, operational and commercial fields are typically not required to think critically about certain design parameters at the design stage. They may not value the scope of simulations in assessing and comparing different clinical trial 'realities'. They may not be fully aware of the impact of some implementation or logistical aspects of conducting a clinical trial on the performance of a given study design like in the case of speed of enrollment and timing of interim analyses, for example.

On the side of challenges, developing simulation code or understanding commercial software to run simulations requires time. The time required to run simulations (computing time may not be easily ignored), summarize results and put together a succinct report add to the working time of the statistician. At the end of the simulation study, there is also the challenge to explain the results in a clear manner, make recommendations and spend time to collect feedback and possibly re-run code with the necessary changes.

Hence, the new educational role for statisticians in addition to conducting and interpreting simulations may appear a significant burden to the profession of statistics in clinical trials. But it is the view of the authors that this is instead a new opportunity for statisticians to become leaders within the clinical team. Simulations offer the

advantage of generating a much more extensive picture of what the trial is, what could happen in different settings, what the potential risks are in terms of decisions and operational consequences. Not only the trial features can be better understood but also the data simulated at patient level can help the researcher to compare designs in setting that are close to reality that may or may not meet the assumptions of the data generating model. All in all, there is much more time required for the statistician to communicate with the team and for the team to understand more complicated scenario analysis than just comparing power levels for different sample sizes.

4.7 Conclusions

In a typical old-fashion setting, statisticians are asked "What sample size do I need to have a Type 1 error rate of 0.05 and power of 90% when the true treatment difference is X with a given standard deviation?". If statisticians take this simple question, that can easily be answered with existing software, as an opportunity to begin a scientific dialogue within the project team, simulations become a tool to answer much more meaningful questions and gain valuable and deeper insight into the best approach(s) for developing a new therapy. Simulations could be a vehicle to develop and strengthen the leadership role of statisticians as important partners to other functions in drug development.

Given the richness of new statistical methodologies and computing technological advances, the utilization of adaptive designs in the pharmaceutical and device industry cannot ignore the important tool of simulations. Not only the drug development paradigm has to shift, but the community of statisticians must also embrace a change in mindset to help develop clinical trials designs that are more efficient and better equipped to make faster and more precise decisions based on interim and accumulating data. This can be achieved with the routine and extensive use of simulations in drug development.

References

Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls. *Statistics in Medicine, 35,* 325–347.

DiMasi, J. A., Feldman, L., Seckler, A., & Wilson, A. (2010). Trends in risks associated with new drug development: Success rates for investigational drugs. *Clinical Pharmacology and Therapeutics, 87*(3), 272–277.

European Medicines Agency. Innovative Drug Development Approaches (2007). Accessed June 14, 2017. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2009/10/WC500004913.pdf.

European Medicines Agency. Report on the implementation of the EMA/CHMP thinktank Recommendations, Areas addressed and recommendations on new emerging issues (2011). Accessed June 14, 2017 http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/09/WC500113212.pdf.

European Medicines Agency. Road map to 2015 (2010). Accessed June 14, 2017 http://www.ema.europa.eu/docs/en_GB/document_library/Report/2011/01/WC500101373.pdf.

FDA Guidance. (27 July, 2016). Adaptive designs for medical device clinical studies.

FDA. (2004). Innovation/Stagnation: Challenge and opportunity on the critical path to new medical products. Accessed June 14, 2017 https://www.fda.gov/downloads/scienceresearch/specialtopics/criticalpathinitiative/criticalpathopportunitiesreports/ucm113411.pdf.

FDA. (2006). Innovation/Stagnation: Critical path opportunities list. Accessed June 14, 2017 https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM077258.pdf.

Gallo, P., et al. (2006). Adaptive designs in clinical drug development—An executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics, 16,* 275–283.

He, W., Pinheiro, J., & Kutznetsova, O. (Eds.). (2014). *Practical considerations for adaptive trial design and implementation*. New York: Springer Science and Business Media.

Kannt, A., & Wieland, T. (2016). Managing risks in drug discovery: Reproducibility of published findings. *Naunyn-Schmiedeberg's Arch Pharmacol, 389,* 353–360.

Paul, S., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., et al. (2010). How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery, 9*(3), 203–214.

Ruberg, S. (2016). Making what's advanced today routine tomorrow. *Statistics in Biopharmaceutical Research, 26*(1), 55–70.

Wang, S. J. (2009). Commentary on "Experiences in model/simulation for early phase or late phase study planning aimed to learn key design elements". *Statistics in Biopharmaceutical Research, 1*(4), 462–467.

Westfall, P. H., Tsai, K., Ogenstad, S., Tomoiaga, A., Moseley, S., & Lu, Y. (2008). Clinical trials simulation: A statistical approach. *Journal of Biopharmaceutical Statistics, 18*(4), 611–630.

Chapter 5
Designing and Analyzing Recurrent Event Data Trials

Stephan Ogenstad

5.1 Introduction

Recurrent event data analysis is common in clinical trials. Literature reviews indicate that most statistical models used for such data are often based on time to the first event or that events within a subject are considered to be independent. Even when taking into account the dependence of the events within subjects, statistical analyses are mostly done with continuous risk interval models, which may not be appropriate for treatments with sustained effects. Furthermore, results can be biased in cases of a confounding factor implying different risk exposure, e.g. in malaria transmission, if subjects are located at zones showing different environmental factors implying different risk exposures (Sagara et al. 2014). Hence, in many prospective randomized controlled clinical trials, events are recurrent, in the sense that the events involve repeat occurrences of the same or different types of events over time. Typical event data consist of times of occurrences of events and the types of events or states that occur. Frequently, an event may be considered as a transition from one state to another and, therefore, multistate models will often provide a relevant framework for event history data. Event history analysis deals with inference for transition intensities and transition probabilities in multistate models. This includes estimation and hypothesis testing for these quantities and analysis of regression models where these quantities are related to explanatory variables observed for the subjects under study. Multistate models are defined by their transition intensities from which transition probabilities may or may not be derived depending on the modeling assumptions. Multistate models are discussed from several points of view in the books and articles by Andersen and Keiding (2002), Andersen et al. (1993), Blossfeld and Rohwer (1995), Courgeau and Lelièvre (1992), and Hougaard (1999, 2000), and Commenges (1999).

S. Ogenstad (✉)
Statogen Consulting LLC, 1600 Woodfield Creek Drive #215, Wake Forest, NC 27587, USA
e-mail: sogenstad@statogen.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_5

The recurrent events are health indicators that assess disease progression or therapeutic effect when subjects are observed over a period of time. It is clinically meaningful to consider whether the treatment a subject is receiving is expected to impact the first event or subsequent events or both. In other words, does the intervention increase the time to the first event or decrease the event number over the study period? In many therapeutic areas, time to the first event is chosen to be the primary endpoint, but this choice then ignores all events after the first one. It is true that statistical approaches for recurrent event endpoints usually are more complex, with less regulatory experience, though there are a number of indications where these endpoints are used, such as hospitalizations in cardiology, asthma and multiple sclerosis. The recurrent event approaches are usually more statistically efficient as information beyond the first event is used. When the follow-up time may be truncated by competing terminal events, it is possible that a subject's observation times may correlate with the competing terminal events themselves, thus making the observation times difficult to assess.

Flexible parametric models of time to the first event or survival can help us in a number of ways. These types of models allow us to obtain estimates of the baseline survival function and its uncertainty which vary smoothly over time. Prediction of survival probabilities and differences, hazard rate functions, hazard differences and ratios, time-dependent effects of covariates, and excess mortality rates in the context of relative survival are just some of the possible outputs from these models.

There are a number of different methods that can be used to evaluate the effects that different treatments can have on subjects in a controlled clinical trial. A few of the methods are to study the 'time to the first event' for the subject, the 'number of events' observed for the subject during the time period that the subject is observed, marginal models, special models with time-dependent covariates, and frailty (random-effects) models. Random effects models are interesting, and our understanding of how they work when applied is beginning to mature. Marginal models are relatively simple to use, interpretable, and flexible, but all of them have limitations. Usually, these models can be fit with standard software such as SAS, Stata and the R package.

In this chapter, we will initially spend some efforts on survival models, since these models form a foundation of many recurrent event models. We are only considering right-censored observations, that is when subjects are still alive at the end of the study and we only have incomplete survival time observations. A crucial problem is whether the available incomplete data enables us to make valid inference on parameters in the multistate model for the complete data. The condition for this is known as independent right-censoring and the interpretation is that a sample observed after independent right-censoring is '*representative*' of the population without censoring. This means that subjects who are censored should have neither lower nor higher risk of future events than subjects who are not censored. We will not cover events that affect trial conduct, such as treatment switching after an event has occurred.

5.2 Methods

5.2.1 Time to First Event

Let T denote a continuous non-negative random variable representing survival time, with probability density function $f(t)$ and cumulative distribution function $F(t) = \mathrm{P}(T \leq t)$. The survival function $S(t) = \mathrm{P}(T > t) = 1 - F(t)$ expresses the probability of a subject being alive at time t. The hazard rate function $\alpha(t) = f(t)/S(t)$ describes the conditional probability of an event occurring at time t, given that the event has not yet occurred. Models based on the hazard rate function can assess whether covariates have an effect on the hazard. If we let $\Lambda(t) = \int_0^t \alpha(u)du$ denote the cumulative or integrated hazard rate function then the survival function can be expressed as $S(t) = \exp(-\Lambda(t))$.

The simplest multistate model is a two-state model where a subject can transition from being '*alive*' (in state 0) to the absorbing state of being '*dead*' (in state 1). Sometimes what is happening to a subject is being viewed as being part of a Markov process. The time it takes until this '*absorbing state*' is reached (or the observational period is censored) is the '*survival time*'. The survival time for a subject will here in the most simple form consist of a random variable, say T, representing the time from a given origin (time 0) to the occurrence of the event '*death*' or we have the knowledge that the observational period is censored. It is seen that $S(t)$ and $F(t)$, respectively, correspond to the probabilities of being in state 0 or 1 at time t. If every subject is assumed to be in state 0 at time 0 then $F(t)$ is also the transition probability from state 0 to state 1 for the time interval from 0 to t. In continuous time the distribution of T may also be characterized by the hazard rate function transition probability from state 0 to state 1 for the time interval from 0 to t. The hazard rate function may be characterized by

$$\alpha(t) = -d \log S(t)/dt = \lim_{dt \to 0} \frac{P(T \leq t + dt | T \geq t)}{dt}$$

that is,

$$S(t) = \exp\left(- \int_0^t \alpha(u)du \right)$$

Thus, $\alpha(.)$ is the transition hazard rate from state 0 to state 1, i.e., the instantaneous probability per time unit of going from state 0 to state 1.

The survival function is often estimated with the Kaplan-Meier (KM) curve (Aalen et al. 2008). It is the most frequently used tool to describe what happened to the subjects in each treatment group. From censored survival data we can easily estimate a survival function by the KM estimator. Figure 5.1 shows KM survival curve estimates

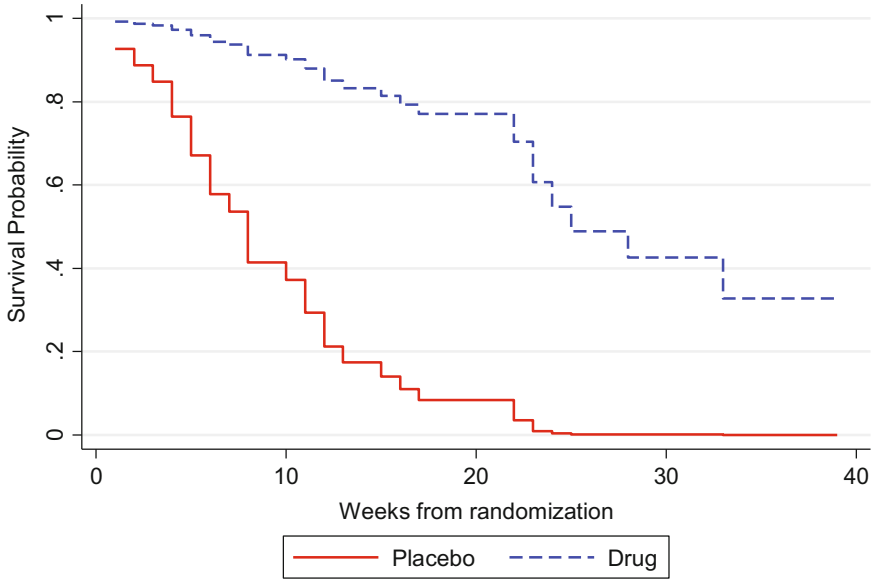


Fig. 5.1 KM survival curve estimates for time to death from cancer for two treatment groups Placebo and Drug

for time to death from cancer for two treatment groups Placebo and Drug. The cancer dataset that ships with the software Stata (cancer.dta) is used, but is entirely fictional.

The least precise parts of the KM curves get the most visual focus, i.e., the right-hand parts of the curves towards the end of the study, where the fewest number of subjects are at risk of the event of death. This is a general criticism of KM survival curve estimates. Kaplan-Meier-type estimates are composed of a sequence of point estimates of the survival functions that are highly serially correlated. Accordingly, KM plots tend to display '*runs*' of values that move away from and back toward the general trend, giving an undulating appearance. This may make the curves difficult to interpret and may lead to the overemphasis of local features (Royston and Lambert 2011).

The estimation of a hazard rate function is more difficult. What can easily be done is to estimate the cumulative hazard rate function $\Lambda(t) = \int_0^t \alpha(u)du$ using the Nelson-Aalen estimator. Figure 5.2 shows the Nelson-Aalen estimates for the same two treatment groups Placebo and Drug described previously.

If the increments of a Nelson-Aalen estimate are smoothed then the new estimates may be used to provide estimates of the hazard rate function themselves. Below are estimates of the hazard rate functions after smoothing of the Nelson-Aalen estimates for the two treatment groups Placebo and Drug (Fig. 5.3).

The smoothing options will, of course, affect the shape of the hazard estimates. We will later on in this chapter show alternative ways of estimating the survival, cumulative hazard and hazard rate functions.

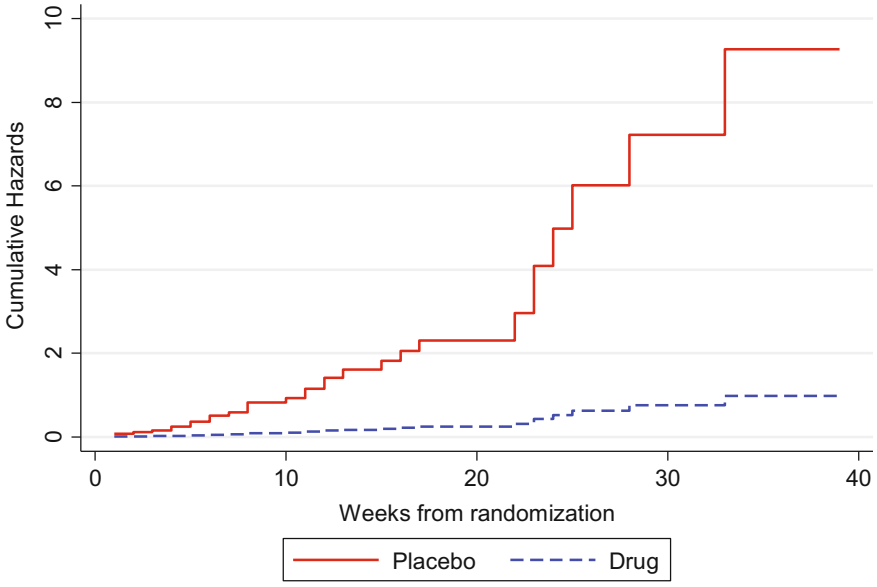


Fig. 5.2 Nelson-Aalen estimates for time to death from cancer for two treatment groups Placebo and Drug

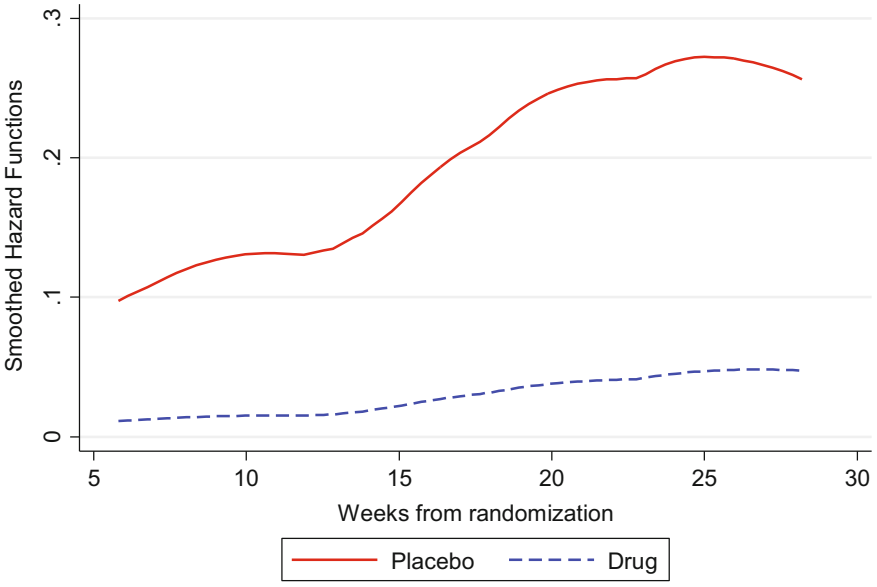


Fig. 5.3 Estimates of the hazard rate functions based on smoothed Nelson-Aalen estimates for time to death from cancer for two treatment groups Placebo and Drug

5.2.1.1 The Cox Proportional Hazards Model

Modeling of censored survival data has since the 1970's almost always been done by the use of the Cox proportional-hazards regression model. The model is in its original form semi-parametric. The hazard rate function for the Cox proportional hazard model (Cox 1972) has the form

$$\alpha(t|\mathbf{z}_i) = \rho_0(t)\exp(\beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_p z_{ip}) = \rho_0(t)\exp(\mathbf{z}_i'\boldsymbol{\beta})$$

which gives the hazard rate at time t for subject i with covariate vector $\boldsymbol{z}_i$ and parameter vector $\boldsymbol{\beta}$. The baseline hazard $\rho_0(t)$ is arbitrary, which in one sense is scientifically comforting, though the function does not extrapolate any information beyond that. An underlying assumption of the Cox model is that the estimated parameters are not associated with time.

Ignoring ties at the moment and conditioning on the existence of a unique event at some particular time t the probability that the event occurs in subject i for which $C_i = 1$ (uncensored) and $T_i = t$ is

$$L_i(\boldsymbol{\beta}) = \frac{\theta_i}{\sum_{j:T_j \geq T_i} \theta_j}$$

where $\theta_j = \exp(z_j'\boldsymbol{\beta})$. Treating the subjects' events as if they were statistically independent, the joint probability of all realized events conditioned upon the existence of events at those times is the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{i:C_i=1} \frac{\theta_i}{\sum_{j:T_j \geq T_i} \theta_j}$$

Its log partial likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i:C_i=1} \left(\mathbf{z}_i\boldsymbol{\beta} - \log \sum_{j:T_j \geq T_i} \theta_j \right)$$

This function can be maximized over $\boldsymbol{\beta}$ to produce maximum partial likelihood estimates of the model parameters.

Several approaches have been proposed to handle situations in which there are ties in the time data. The partial likelihood for recurrent failure times is the case when two or more subjects are recorded as dying at the same time. Breslow (1975) developed a method that is the default for many statistical software packages, but it is not the default for the R package. Breslow's method uses the partial likelihood, expressed as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{I} \frac{\prod_{j \in D(t_{(i)})} \phi_j}{\left(\sum_{j \in R(t_{(i)})} \phi_j\right)^{|D(t_{(i)})|}}$$

where $\left|D(t_{(i)})\right|$ is the number of subjects that fail at time $t_{(i)}$.

Breslow's method describes the approach in which the procedure described above is used unmodified, even when ties are present. An alternative approach that is considered to give better results is Efron's method (Efron 1974). The Cox model may be specialized if a reason exists to assume that the baseline hazard follows a particular form. In this case, the baseline hazard $\rho_0(t)$ is replaced by that particular function. An alternative to Cox's model is the additive regression model due to Aalen (Aalen et al. 2008), which assumes that the hazard rate of a subject i with p covariates z_{i1}, …, z_{ip} takes the form

$$\alpha(t|\mathbf{z}_i) = \beta_0(t) + \beta_1(t)z_{i1} + \ldots + \beta_p(t)z_{ip}.$$

For this model $\beta_0(t)$ is the baseline hazard, while the *regression functions* $\beta_j(t)$ describe how the covariates affect the hazard rate at time t. For the Cox and additive regression model hazard rate functions, the covariates are assumed to be fixed over time. More generally, one may consider covariates that vary over time (Aalen et al. 2008). The generic term parametric proportional hazards models can be used to describe proportional hazards models in which the hazard rate function is specified. The Cox proportional hazards model is sometimes called a semiparametric model by contrast.

The R package uses Efron's partial likelihood, as it is considered a closer approximation to the exact partial likelihood. Efron's partial likelihood has the following shape

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{I} \frac{\prod_{j \in D(t_{(i)})} \phi_j}{\prod_{k=1}^{|D(t_{(i)})|} \left(\sum_{j \in R(t_{(i)})} \phi_j - \frac{k-1}{|D(t_{(i)})|} \sum_{j \in D(t_{(i)})} \phi_j\right)}$$

An extension of the proportional hazards model is to allow for multiple strata in the fitting procedure. That is, we assume that the subjects can be broken into multiple groups, and the hazard rate function for subjects in the kth group is

$$\rho_{0k}(t) \exp(\mathbf{z}_i' \boldsymbol{\beta}).$$

A common use of stratification is in multicenter trials. Because of different subject populations and referral patterns, different centers in the trial may have quite different hazard rates, yet a common treatment effect across centers. In this way, strata play a similar role to multiple intercept terms in an analysis of covariance model. Each baseline hazard captures the baseline rate for an event. When events are of different types, we have in reality different baselines. If we, for instance, are studying heart

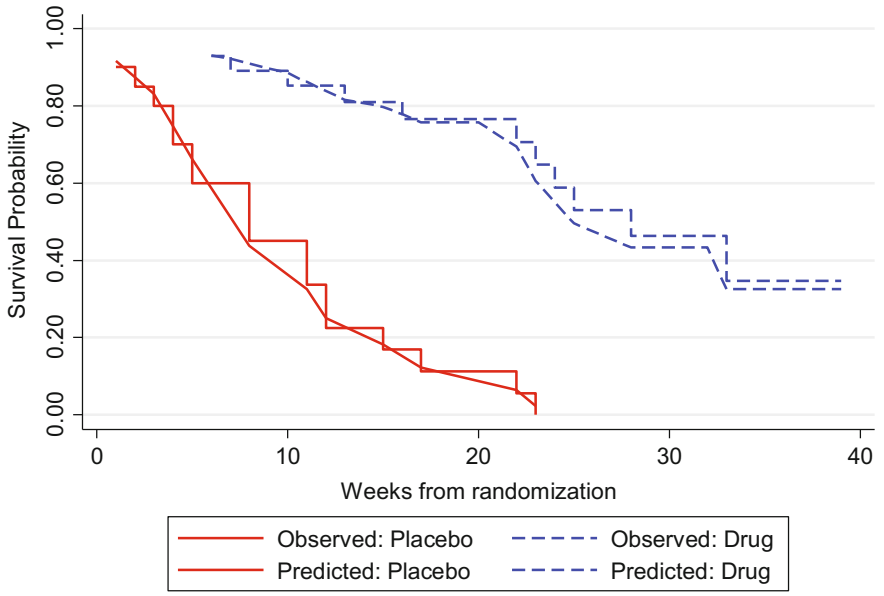


Fig. 5.4 KM and Cox model estimates for time to death from cancer for two treatment groups Placebo and Drug

attacks, are the first and second heart attacks the same type of event? Well, we would only know if we investigate it.

The Kaplan-Meier and Cox estimation provide estimates of the survival functions. Continuing to use our cancer data, Fig. 5.4 displays the KM and Cox model estimates.

5.2.1.2 Extending the Cox Model for the Two-State Case

The main purpose of the Cox model in its simplest form is to estimate hazard rates assuming that the hazards are proportional to each other. Because the model can be embedded in a counting process framework (Andersen et al. 1997), the model can be extended in many different ways to answer questions across a wide range of situations, where we need to obtain informative estimates of quantities that include hazard rates and their differences and ratios, survival curves and their differences, rates, and survival at given time points. By '*informative*' we mean unbiased estimates that are smooth functions.

Parametric survival models generally provide smooth estimates of the hazard and survival functions for any combination of covariate values. The exponential model is often used when planning a clinical trial and for calculating the power and sample sizes. Though, the exponential survival model is a rather unrealistic model since it is assumed that the hazard rate function is constant over the whole observational study period. This model can be generalized by splitting the observational period into intervals. The choice of the number of intervals and where to place the cutpoints

is of course subjective. With the piecewise exponential model, the time scale is split into several intervals, where we assume that the hazard rate function is constant within each interval but can vary from interval to interval. The hazard rate function for the piecewise exponential model can be written $h_{ij}(t, \mathbf{z}_i) = \alpha_j \exp(\mathbf{z}_i'\boldsymbol{\beta})$, where the subscript i is for subject and j is for the interval. Modeling the data with the Poisson approach allows us to think about survival time in a different way from that in standard survival analysis. Usually, survival time is considered to be the outcome variable and we have to use special methods to account for the censoring process. With the Poisson approach, it becomes clearer that we are modeling rates. We have a binary variable as an outcome, and our models investigate variation in the corresponding rates. There are many factors that cause systematic variation in rates, for example, age and gender, but also time. In the Poisson framework, we can, therefore, consider time to be a covariate, as opposed to a response. Thus we can adjust for time just as we would for any other covariate. Time-dependent effects of a covariate of interest are then simply an interaction between time and the covariate. The Poisson model with a split at each unique failure time gives us the Cox model. However, we do not want to fit a model with so many parameters. An important question is, what is the effect of changing the number of time intervals of the parameters of interest (usually log hazard ratios)? The problem with the piecewise exponential model is that if we choose too few intervals we may miss important changes in the hazard rate; if we choose too many, we end up with too many parameters, and the underlying shape of the hazard rate is difficult to see because of random variation.

5.2.1.3 Royston-Parmar Models

The use of parametric models for the type of data we so far have considered may have some advantages. The non-proportional hazards that are a potential difficulty with the Cox model, could sometimes be handled in a simpler way, and the visualization of the hazard rate function could be much easier. Royston–Parmar (RP) models (Royston and Parmar 2002, Lambert and Royston 2009) have great flexibility with respect to the shapes of the survival distributions they can model. Familiar standard parametric survival models are the starting point for the generalizations called RP models. Weibull, log logistic, and lognormal models can be generalized to proportional hazards, proportional odds, and probit-scaled RP models, respectively. The additional flexibility of RP models arises because the baseline distribution function is represented as a restricted cubic spline function of log time instead of simply as a linear function of log time. Modeling with spline functions generates some additional complexity. The additional complexity is determined by the number and the positions of the connection points in log time, known as *knots*, of the spline's cubic polynomial segments. Estimation of parameters is by maximum likelihood. Quite often, the characteristics of the fitted model are rather insensitive to the number and particularly the position of the knots, lending a certain robustness to the process of model selection. The restriction that the transformed survival function be linear in Ln(t) is, in practice, severely limiting and is not really necessary. In RP models, we may relax linearity and allow nonlinear functions. There are many possible fami-

lies of nonlinear functions that we could use. Because cubic splines are flexible yet relatively simple to work with and understand, Royston and Parmar (2002) chose them as their preferred tool to extend standard models. The result is a major advancement in the practical usefulness of parametric survival analysis and in the range of applications that can be tackled.

In cancer survival trials, one often wants to know the impact of covariates on the mortality rate for a particular cancer diagnosis. Since cancer is mostly a disease of old age, many people may die of diseases other than the specific type of cancer they were originally diagnosed with. Relative survival is a measure of patient survival corrected for the effect of other causes of death by utilizing the patients' expected survival. Both Poisson models and Royston–Parmar (RP) models can be extended to relative survival by incorporating information on expected survival or mortality. Relative survival is related to the concept of competing risks (Gamel and Vogel 2001). We there assume that an individual is at risk of either dying of their cancer or dying of another cause. In relative survival models, we can deal with this issue by incorporating expected mortality, which can usually be obtained from routine data sources. Traditionally, simple piecewise models have been used for relative survival, but all the advantages of standard parametric survival models also apply to relative survival models.

The baseline survival function in a Cox model is available only in the estimation sample. To predict survival outside the estimation sample, we need special measures, such as interpolation or even extrapolation. Using special measures limits the applications of the Cox model in some situations. An important case arises when we wish to validate a survival model in an independent sample, a task that necessitates out-of-sample prediction. There are at least two situations in which this is useful. One is by interpolating or extrapolating the baseline or other survival functions at time points not represented in the estimation sample. The other is by predicting survival probabilities or other quantities of interest from a model on a derivation sample onto individuals in an evaluation sample (that is, external validation). Interpolation is helpful, for instance, when we wish to plot a survival function for an individual, a group, or a covariate pattern as a smooth curve at a suitable choice of time points within the range of the observed follow-up time. We need the extrapolation when we want to project a modeled survival function into the future. Successful external validation is usually regarded as the gold standard of potential usefulness of a proposed prognostic model (Altman and Royston 2000).

The Hazard rate function is of utmost relevance in clinical medicine since it is a decidedly meaningful measure of disease course, and is the basis against which relative hazard effects are estimated. Fuchs et al. (1994) report on a double-blind randomized multicenter clinical trial designed to assess the effect of rhDNase (purified recombinant form of the human enzyme DNase I) versus placebo on the occurrence of respiratory exacerbations among patients with cystic fibrosis. The subjects in these treatment groups are susceptible to an accumulation of mucus in the lungs, which leads to pulmonary exacerbations and deterioration of lung function. The occurrences of exacerbations over the study period were recorded for each subject. The estimated hazard functions in the Fig. 5.5 are derived from the Cox model and the

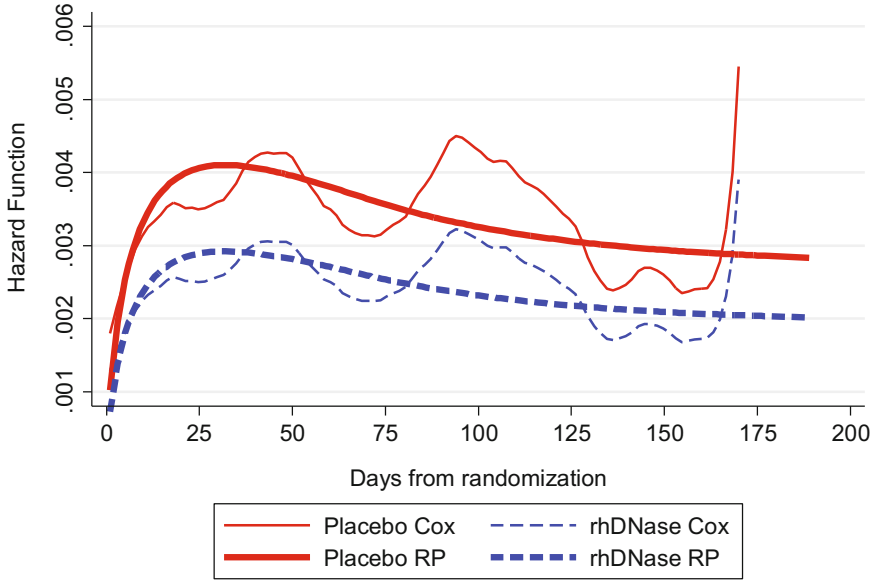


Fig. 5.5 Cox model and the Royston-Parmar model estimates under the proportional hazards assumption from rhDNase data

Royston-Parmar model under the proportional hazards assumption. The thicker pair of lines show estimates of the hazard rate functions from the Royston-Parmar model and the thinner lines from the Cox model, using kernel smoothing, in the Placebo and rhDNase groups, respectively.

The Royston-Parmar model gives a more plausible trajectory of the hazard for the patients, then the rugged course that is shown from the estimates based on the Cox model. In the RP model, the hazards seem to be highest about one month after randomization and decreases after that time. The hazards are substantially reduced by the rhDNase treatment. The proportional hazards condition forces the curves to be proportional to each other. Even after 175 days the hazard in the rhDNase treatment arm is still substantial but reduced by about one third. The fact that the curve does not approach zero suggests that the disease is chronic. We have obtained quite a lot of useful information. Even if we relax the proportional hazards assumption, the plot of the ensuing hazard rate functions (not shown) are very similar to the thick lines in the figure. So, our conclusion about the treatment effect seems to be robust.

The baseline hazard contains useful information. If we are told that the mortality rate is double for subjects with a particular exposure, then we want to know what reference value this doubling refers to. In a survival model, the reference is usually the baseline hazard rate, which usually changes as a function of time. Thus even if the proportional hazards assumption is reasonable, the impact of a particular exposure in absolute terms depends on how long time has passed since the time origin (diagnosis, randomization, start of treatment, etc.) and the magnitude of the underlying hazard

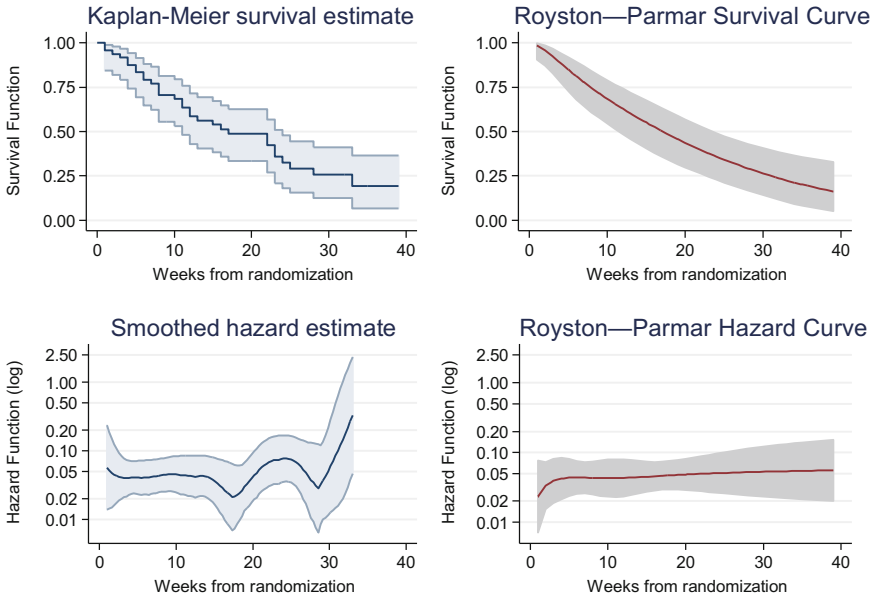


Fig. 5.6 Estimated survival and hazard rate functions with confidence intervals, for time to death from cancer for two treatment groups Placebo and Drug

rate. Flexible parametric survival models can help us in a number of ways. For instance, these models allow us to obtain an estimate of the baseline survival function and its uncertainty which vary smoothly over time.

We will illustrate (Fig. 5.6) the survival distributions and hazard functions using nonparametric techniques (Kaplan–Meier and smoothed hazard functions, respectively) and a flexible parametric technique (Royston–Parmar models) using the cancer dataset.

The survival curves indicate a median time to event of about 16–20 weeks. The Kaplan–Meier curve shows a slight downturn after about 22 weeks, which is not reflected in the survival curve from the Royston-Parmar estimates. The smoothed nonparametric hazard estimate shows a corresponding upturn about 30 weeks. Whether the feature is "real" or not is questionable—it seems surprising that the event rate would start to increase after 18 weeks and then gradually turning down at 30 weeks, and then again from thereon shoot up. The pointwise confidence intervals (CIs) from the smoothed hazard estimate are wider than that from the Royston-Parmar. Conditional on a parsimonious parametric model, CIs are generally too narrow because they do not take model uncertainty into account. Nonparametric CIs make fewer assumptions and tend to be wider. Also, they are implicitly high-dimensional and noisy.

The Royston-Parmar models can equally be used to perform multistate survival analysis (Crowther 2016).

5.2.2 Multiple Events Per Subject

5.2.2.1 Multistate Models

We have considered situations where each subject can only have one event. If death is the outcome, then clearly it is not possible to have more than one event. However, if the event is the recurrence of disease or readmission to hospital, then it is possible for each subject to have more than one event. As a continuation of survival analysis, we will consider another type of multivariate data in the setting of competing risks, where $T_1, T_2, \ldots, T_k$ represent survival times to different causes of death. Estimation of these models is complicated by the fact that we only observe $T = \min\{T_1, \ldots, T_k\}$ where even T can be censored, so that none of the k events have occurred. Yet, another type of multivariate data involves transitions among several types of states, where some of them might be terminal, but not all. This combines elements of competing risk models with models for series of events.

The framework for these types of models can be set up in the following way: Suppose there is a total of m subjects accrued into a study and each subject is at risk for a particular type of recurrent event. Let $(0,\tau]$ represent the period of observation and let $N_i^*(u)$ be a right-continuous integer function representing the number of events experienced by subject i over the interval $(0, u]$, $i = 1, 2,\ldots, m$, $0 < u \leq \tau$. During the observation period $(0, \tau]$, some subjects may experience an event which terminates their recurrent event processes (e.g. death), but subjects may also withdraw from the study according to some random censoring mechanism which is independent of the recurrent event and terminal event processes. For $i = 1, \ldots, m$, let T_i be the time of the terminating event, C_i the censoring time, $X_i = \min(T_i, C_i)$, $\pi(t) = \mathrm{P}(X_i \geq t)$, and $\delta_i = I(X_i = \mathrm{d}T_i)$, where $I(\cdot)$ is an indicator function. We let $N_i(t) = N_i^*\{\min(t, T_i)\}$ denote the number of recurrent events observed over $(0, t]$ in the presence of death. The data contributed by each subject then take the form $(\{N_i(u), 0 < u \leq X_i\}, X_i, \delta_i)$, $i = 1,\ldots, m$. Let $Y_i(t) = I(X_i > t)$ be the at risk indicator function which is one when subject i is under observation and at risk for an event at time t and is zero otherwise. We suppose initially that we have a single sample of subjects.

The most important class of models is the continuous time Markov process $X(t)$ on the finite state space $S = \{1,\ldots, p\}$ where the dependence of transition hazard rate function $\alpha_{nj}^i(t)$ on the history X_t is only through the current state of $X(t)$ and possibly via time-fixed covariates. Statistical models are usually obtained by specifying the class of transition intensities $(\alpha_{nj}^i(t))$ for each subject i.

The most important deviations from the Markov property in practice are various kinds of duration dependence, where transition intensities depend on other time origins than $t = 0$, typically the time of entry to the present state. There are two main approaches to handling these. As long as transition intensities depend only on one-time origin each (for example, all intensities depend only on duration in the present state), a model for the multistate process may be obtained by combining independent submodels for each transition hazard rate. These may, in turn, be modeled as constant or piecewise constant or by non- or semiparametric models, and as long as

there is a unidirectional flow in the model, transition probabilities are still straightforward explicit functionals, which may be estimated by plugging in the hazard rate estimates. Variance calculations may, however, become less direct (Andersen and Keiding 2002).

5.2.2.2 Univariate Recurrent Events

At the moment, we are only concerned with univariate events, i.e., events of the same kind. Because the various events occur to the same subject, the waiting times will in general not be independent. Since the events occur one after the other, it will generally be the case that only the last interval can be censored.

With recurrent events, we can expect a correlation between the times to event of a given subject. For instance, subjects with severe disease will tend to have more events and a shorter time between events than those with mild disease. The most common models used are (i) Generalized estimating equations model using a Poisson or Negative Binomial distribution, and three extended Cox models: (ii) the Andersen-Gill counting process (AG) (Andersen and Gill 1982), (iii) the Prentice-Williams-Peterson counting process (PWP) (Prentice et al. 1981), (iv) Wei, Lin, and Weissfeld (WLW) (Wei et al. 1989; Lin 1994) and (v) the frailty model (Gutierrez 2002). For the marginal models, the correlation is dealt with using a robust sandwich-based estimator to avoid inflation of type I error due to multiple observations per subject which do not require specification of the correlation matrix (Kelly and Lim 2000). Consideration needs to be taken whether the events are ordered or not. Ordered events could be, for instance, first, second, third, … hospitalization. Unordered events could, for instance, be of different types, such as '*hospitalization*', '*withdrawal*', and '*death*', where '*death*' is a competing event. For more details of the approaches, see Therneau (1997) and Therneau and Grambsch (2000). One can fit similar models within the Royston-Parmar framework (Royston and Parmar 2002; Lambert and Royston 2009).

5.2.3 Poisson Regression

A Poisson process can be described via the hazard rate function that is of the form

$$\alpha(t|H(t)) = \rho(t) \quad t > 0,$$

where $\rho(t)$ is a nonnegative integrable function. It is also assumed that the cumulative hazard rate

$$\mu(t) = \int_0^t \rho(u)du \quad t > 0,$$

is continuous and finite for all $t > 0$. It is seen from the hazard rate function above that the Poisson process is Markovian. The probability of an event in $(t, t + \Delta t)$ may depend on t but is independent of the history $H(t)$.

Poisson regression is a generalized linear form of regression analysis used to model count response data. Poisson regression assumes the response variable Y has a Poisson distribution and assumes that the logarithm of its expected value can be modeled by a linear combination of unknown parameters. The Poisson regression model is frequently used to analyze count data when the dependent variable represents the number of independent events that occur during a fixed period of time (Prentice et al. 1981, Sagara et al. 2014). The method assumes that all events are independent and is based on event rates, where the total number of events is divided by the follow-up time. The conditional mean of Y (the number of events) can be written as:

$$Ln(Y|\mathbf{Z}, \boldsymbol{\beta}) = \mathbf{Z}_i\boldsymbol{\beta}$$

where $\mathbf{Z}_i\boldsymbol{\beta} = \beta_0 + \beta_1 Z_1 + \ldots + \beta_k Z_k$ of k parameters and Ln is the natural logarithm function.

The probability function for a unit-time interval for a subject i can be expressed as

$$f_Y(y_i; \mu_i) = e^{-\mu_i}\mu_i^{y_i}/y_i!$$

for $y = (0, 1, \ldots)$ and $\mu_i > 0$. The mean and variance are both equal to μ_i. With subscripts indicating subject i's observation the log-likelihood function can be written as

$$L(\mu_i; y_i) = \sum [y_i \ln(\mu_i) - \mu_i - \ln(y_i!)]$$

The parameter μ_i can be reparameterized as $\exp(\mathbf{z}_i'\boldsymbol{\beta})$, and therefore the log-likelihood function can be written as

$$L(\boldsymbol{\beta}_i; y_i) = \sum [y_i(\mathbf{z}_i\boldsymbol{\beta}) - \exp(\mathbf{z}_i\boldsymbol{\beta}) - \ln(y_i!)]$$

5.2.4 Negative Binomial Regression

One of the key features of the Poisson distribution is that the variance equals the mean. However, one often finds that overdispersion is frequent in count data. Overdispersion in a Poisson model occurs when the variance of the response is greater than the mean. One approach to handling the overdispersion is to add covariates to the model. Though, even after conditioning on covariates, there could still be more inter-subject variation in event occurrence than accounted for by a Poisson process. Another approach is then to model the overdispersion by adding a multiplicative random

effect to represent unobserved heterogeneity. Doing so will lead to the negative binomial regression model where the conditional distribution of the outcome Y, given an unobserved variable θ, is indeed Poisson with mean and variance $\theta\mu$. The variable θ captures unobserved factors that increase (if $\theta > 1$) or decrease (if $\theta < 1$) relative to what we would expect given the observed values of the covariates. In this model, the data would be Poisson if only we could observe θ. Unfortunately, we do not. Instead, we make an assumption regarding its distribution and integrate θ out of the likelihood, effectively computing the unconditional distribution of the outcome. It is mathematically convenient to assume that θ follows a gamma distribution. The unconditional distribution of the outcome is the negative binomial distribution (Cook and Lawless 2007; Hilbe 2007).

5.2.5 *Extended Cox Models for Recurrent Events*

As we mentioned, recurrent event data are correlated since multiple events may occur within the same subject. While using frailty models is one method to account for the correlation in recurrent event analyses, a simpler approach that can also account for this correlation is the use of robust standard errors (SEs). With the addition of robust SEs, recurrent event analysis can be done as a simple extension of either semi-parametric or parametric models.

If interest focuses on recurrent occurrences of a given event, for instance, hospitalization, then another model than the Cox model should be considered. In applications of such a model, an interesting functional is often the expected number of occurrences of the event over the time interval $(0, t]$. The corresponding semi-parametric estimate of the cumulative expected number of events over $(0, t]$ for subject i is

$$\hat{E}[N_i(t)] = \int_0^t \hat{\rho}_0(s) \exp(\mathbf{z}_i' \hat{\boldsymbol{\beta}}) ds$$

where $N_i(t)$ is the number of events for subject i over $(0, t]$. This is the same as the generalized Nelson–Aalen, or Breslow, estimate from survival analysis. (Cook and Lawless 2002; Andersen et al. 1993).

Cumulative Sample Mean Function

Plots like the one in Fig. 5.7 have limitations since it is often not easy to determine visually whether a trend or other patterns exist in data.

A visually more informative function is the cumulative sample mean function (Cook and Lawless 2007). The function can be defined as follows. Suppose that m individual processes are observed, with each process being observed over the time interval $(0, t]$. Let $N_i(t)$ represent the number of events over the time interval $(0, t]$ for the ith process. Then the cumulative sample mean function is

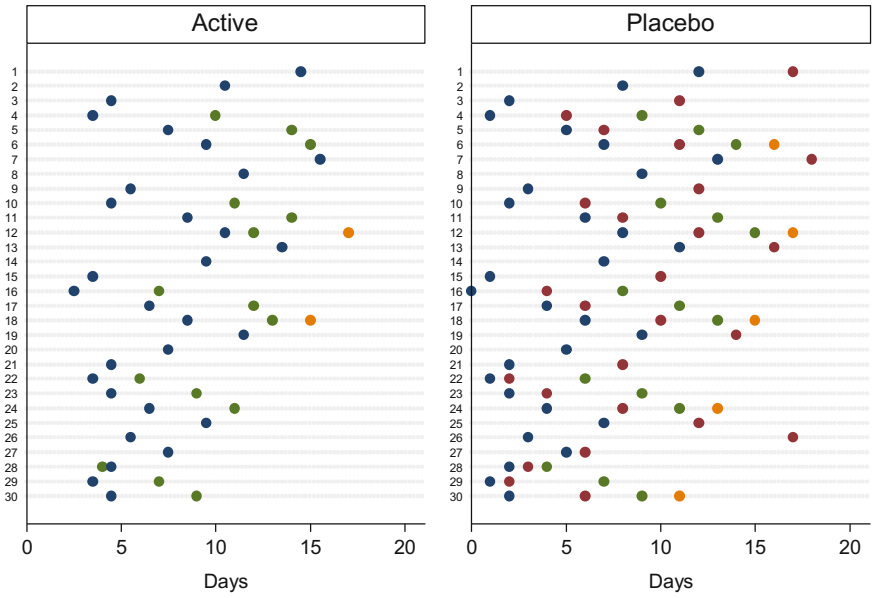


Fig. 5.7 Event plots from time of randomization for tumor occurrence in 60 subjects (30 subjects on Active and Placebo, respectively)

$$\hat{\mu}(t) = \frac{1}{m} \sum_{i=1}^{m} N_i(t).$$

The same data as in Fig. 5.7 is used to display the cumulative sample mean function (Fig. 5.8).

5.2.5.1 The Andersen-Gill Model (AG)

The counting process, or Andersen-Gill, approach to recurrent event modeling assumes that each recurrence is an independent event, and does not take the order or type of event into account. In this model, follow-up time for each subject starts at the beginning of the study and is broken into segments defined by events (recurrences). Subjects contribute to the risk set for an event as long as they are under observation at that time (not censored). The model is simple to fit as a Cox model with the addition of a robust standard error estimator, and hazard ratios are interpreted as the effect of the covariate on the recurrence rate over the follow-up period. This model would be inappropriate, however, if the independence assumption is not reasonable.

External covariates $x(t)$, which include fixed covariates, can be incorporated in a Poisson process by specifying the hazard rate as a function of t and the covariate history $x^{(t)} = \{x(u) : 0 \leq u \leq t\}$. This is usually done by defining covariate vectors

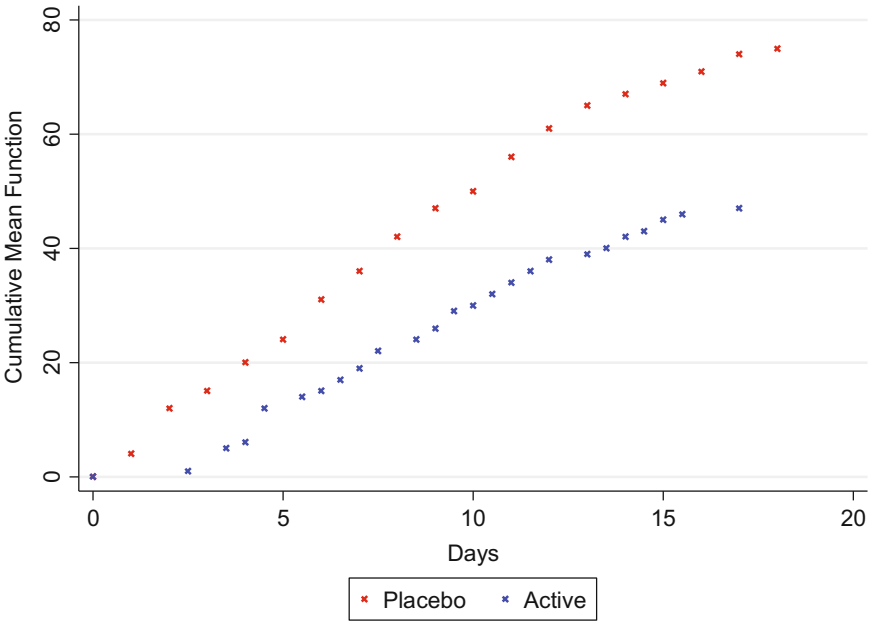


Fig. 5.8 The cumulative sample mean function from time of randomization for tumor occurrence in 60 subjects (30 subjects on Active and Placebo, respectively)

$\mathbf{z}(t)$ that are based on $x^{(t)}$ and then considering the multiplicative intensities of the form

$$\rho(t|x^{(\infty)}) = \rho(t|x^{(t)}) = \rho_0(t)\exp(\mathbf{z}'(t)\boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is a vector of regression parameters of the same length as $\mathbf{z}(t)$. The positive valued function $\rho_0(t)$ is often called the baseline rate or intensity and corresponds to a subject for whom $\mathbf{z}(t)=0$ for all $t>0$. This model is sometimes referred to as a log-linear model. The exponential term can be replaced by a different positive term but has been chosen for mathematical convenience. When the baseline function $\rho_0(t)$ is specified as nonparametric then the model is semiparametric and is called the Andersen-Gill (AG) (1982) model.

The AG model is an extension of the Cox model and uses the counting process timescale for all events. The time-scale does not reset to 0 after an event but continues from the time point of the event. Data for each subject needs to be entered in the counting process style, with a start time, stop time and censoring indicator for each event. The model is close in spirit to Poisson regression and the increments are assumed to be independent. Each gap time (interval from one event to the next) contributes to the likelihood and the model assumes that the events are independent. The AG model splits the time scale where the split points are defined by the time point when the events occur. The time intervals are non-overlapping; that is, the start

time of a new event is the ending time of the preceding event. In the AG model, the underlying shape of the baseline hazard is assumed to be the same for all events; that is, there is no stratification by event number. Although not specified in the original article, cluster-based robust standard errors are usually used.

The sandwich robust standard error of Lin and Wei (1989) which is a variance-correction technique, is usually employed together with these Cox extended models to avoid inflation of type I error due to multiple observations per subject which do not require specification of the correlation matrix.

5.2.5.2 Conditional Counting Process Model by Prentice-Williams-Peterson (PWP)

The PWP model is a conditional model, similar to the AG model, but stratified by events. The hazard rate function is written as:

$$\rho_{ik}(t|x^{(t)}) = \rho_{0k}(t)\exp(\mathbf{z}'_{ik}(t)\boldsymbol{\beta})$$

$\rho_{0k}(t)$ represents the event-specific baseline hazard for the kth event over time. In this model, a subject is assumed not to be at risk for a subsequent event until a current event has terminated. The PWP model is similar to the AG model in that it uses nonoverlapping time intervals (gap times) for each subject. As for the AG model, it is not possible to be at risk of the second event before the first event has occurred. The PWP model differs from the AG model in that the baseline hazard for each event k is allowed to be different; that is, there is stratification by event number.

5.2.5.3 The Wei, Lin, and Weissfeld (WLW) Model

Suppose there are n subjects and each subject can experience up to K potential events. Let $\mathbf{Z}_{ki}(t)$ be the covariate process associated with the kth event for the ith subject. The marginal Cox model is given by

$$\rho_{ik}(t|x^{(t)}) = \rho_{0k}(t)\exp(\mathbf{z}'_{ik}(t)'\boldsymbol{\beta}_k),\ k = 1,\ldots,K; i = 1,\ldots,n$$

$\rho_{0k}(t)$ is the (event-specific) baseline hazard function for the kth event and $\boldsymbol{\beta}_k$ is the (event-specific) column vector of regression coefficients for the kth event. The WLW model estimates $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K$, by the maximum partial likelihood estimates $\hat{\boldsymbol{\beta}}_1,\ldots,\hat{\boldsymbol{\beta}}_K$, respectively, and uses a robust sandwich covariance matrix estimate for $(\hat{\boldsymbol{\beta}}'_1,\ldots,\hat{\boldsymbol{\beta}}'_K)'$ to account for the dependence of the multiple failure times.

The WLW model uses overlapping time intervals for each subject and stratum so that each stratum is fit separately, and then the estimates are combined. This implicitly forces all strata and covariate interactions to be present. This is equivalent to fitting all of the data at once, i.e., the events are occurring in parallel. The model treats an

ordered dataset as though it were an unordered dataset in a competing risks problem. Thus, each event or event type is in its own stratum and all time intervals starting at 0. Hence, the WLW approach considers each event to be a separate process, so subjects are at risk for all events from the start of follow-up, regardless of whether they experienced a prior event. This model is appropriate when the events are thought to result from different underlying processes, so that a subject could experience the 3rd event, for example, without experiencing the 1st. Although this assumption seems implausible with some types of data, like cancer recurrences, it could be used to model injury recurrences over a period of time, when subjects could experience different types of injuries over the time period that have no natural order. There is a need to specify the total number of events in advance. The method also analyzes the gap times between different events.

5.2.5.4 Competing Risks

Traditional survival analysis methods assume that only one type of events of interest occurs. A way to avoid dealing with competing events in a more complex model than the Cox model is to construct composite endpoints. An example of this is when studying cardiovascular outcomes in type 2 diabetes, where the primary composite outcome is the time-to-event of the first occurrence of death from cardiovascular causes, nonfatal (including silent) myocardial infarction, or nonfatal stroke (Marso et al. 2016). Models in which there are different types of events (*multiple destinations*) are also of interest. Competing risks occur when a subject is at risk of more than one type of event, but can actually experience only one of them. The most common case is when the different events are death from different diseases, such as cancer, heart disease, or an infection. Competing risk models are a special case of multistate models in which each of the different events are absorbing states (Andersen et al. 2002). In competing risks, a subject is at risk of dying from one of, say K, different causes, but can only actually die of one cause.

More complex methods exist to allow the investigation of several types of events in the same study, such as death from multiple causes. Competing risks analysis is used for these studies in which the survival duration is ended by the first of several events. Special methods are needed because analyzing the time to each event separately can be biased. Specifically, in this context, the Kaplan-Meier method tends to overestimate the proportion of subjects experiencing events. Competing risk analysis utilizes the cumulative incidence method, in which the overall event probability at any time is the sum of the event-specific probabilities. The models are generally implemented by entering each study participant several times—one per event type. For each study participant, the time to any event is censored on the time at which the patient experienced the first event.

The two most significant measures in competing risks are the cause-specific hazard rate and the cumulative incidence function. The cause-specific hazard rate function for cause k, say $h_k(t)$, gives the hazard rate at time t conditional on not having died of any of the K possible causes of death. The cause-specific hazard, $h_k(t)$, can be

estimated by treating events due to competing causes as censored observations. The K cause-specific hazard rates are usually estimated by fitting K separate models or by stacking the events (having K rows of data per subject) and fitting a model stratified by cause (Lunn and McNeil 1995). The second most important measure is the cumulative incidence function, say $C_k(t)$, for the kth competing event. This gives the probability, as a function of time, that a subject dies of cause k in the presence of competing risks. It recognizes that a subject cannot die of cause k if that subject has already died of one of the competing causes. The cumulative incidence function is also known as the crude probability of death (Tsiatis 2005). It can be contrasted with the net probability of death, which gives the probability of dying in a situation where it is impossible to die of other causes. cumulative incidence functions give probabilities of death where subjects are always at risk of death from several different causes. The cumulative incidence is calculated from a relative survival model and is defined as

$$C_k(t) = \int_0^t h_k(u) \exp\left\{ - \int_0^u \sum_{k=1}^{k} h_k(v) dv \right\} du = \int_0^t h_k(u) \prod_{k=1}^{K} S_k(u) du.$$

$C_k(t)$ can be calculated by using the Stata package (Fine and Gray 1999).

A nonparametric analysis of recurrent events in the presence of death as a competing risk has been developed by Ghosh and Lin (2000) and by Li and Lagakos (1997).

5.2.5.5 Period Analysis

Cancer survival measures the effectiveness of health-care systems. Persistent regional and international differences in survival represent a source of information that may be used to avoid early death. Differences in survival have impelled or steered cancer control strategies. Statistics reflective of patient survival should be as current as possible. The traditional methods for analyzing survival have important shortcomings with regard to how current they are with respect to long-term cumulative survival estimates. An alternative approach denoted '*period analysis*', that may be used to overcome or reduce these constraints. When cancer survival is improving over time, the use of older data underestimates the survival proportion. One potential solution to this is to use period analysis to obtain more up-to-date estimates of patients' survival (Brenner and Gefeller 1997). This approach has become widely established in the analysis of population-based cancer survival. For example, it has been used in a number of recent international comparisons of cancer survival (Coleman et al. 2011; Møller et al. 2010). Period estimates of patient survival are usually calculated separately in subgroups of interest using life table methodology. Up-to-date estimates of patient survival using period analysis are based on artificially truncating individuals' survival times prior to a recent cutoff in calendar time. This has the effect of using

individuals diagnosed in a recent time period for short-term survival and individuals diagnosed further back in time for longer term survival.

In the Coleman et al. (2011) study, data from population-based cancer registries in 12 jurisdictions in six countries were provided for 2.4 million adults diagnosed with primary colorectal, lung, breast (women), or ovarian cancer during 1995–2007, with follow-up to Dec 31, 2007. Data quality control and analyses were done centrally with a common protocol, overseen by external experts. They estimated 1-year and 5-year relative survival, constructing 252 complete life tables to control for background mortality by age, sex, and calendar year. Also, they reported age-specific and age-standardized relative survival at 1 and 5 years, and 5-year survival conditional on survival to the first anniversary of diagnosis. In addition, they examined incidence and mortality trends during 1985–2005. Their findings were that relative survival improved during 1995–2007 for all four cancers in all jurisdictions.

In the Møller et al. (2010) study, several international studies reported that survival from breast cancer is lower in the United Kingdom than in some other European countries. They compared breast cancer survival between the national populations of England, Norway, and Sweden, with a view to identifying subsets of patients with particularly good or adverse survival outcomes. They also extracted cases of breast cancer in women diagnosed 1996–2004 from the national cancer registries of the 3 countries. The study comprised 303,657 English cases, 24,919 Norwegian cases and 57,512 cases from Sweden. Follow-up was in 2001–2004. The main outcome measures were 5-year cumulative relative survival and excess death rates, stratified by age and period of follow-up.

5.2.5.6 Frailty Models

Correlated survival data can arise due to recurrent events experienced by an individual or when observations are clustered into groups. Either due to lack of information or for feasibility, some covariates related to the event of interest may not be measured. Frailty models account for the heterogeneity caused by unmeasured covariates by adding random effects that act multiplicatively on the hazard function. Frailty models are essentially extensions of the Cox model with the addition of random effects. Although there are various classification schemes and designation used to describe these models, four common types of frailty models include shared, nested, joint, and additive frailty.

The frailty model, introduced in the biostatistical literature by Vaupel et al. (1979), and discussed in detail by Hougaard (1984, 1986a, b, 1995), Duchateau and Janssen (2008), and Wienke et al. (2001), accounts for the heterogeneity in baseline. This model is an extension of the proportional hazards model in which the hazard rate function depends upon an unobservable random variable. Subjects may be exposed to different risk levels, even after controlling for known risk factors, because of some relevant unobserved covariates. In a shared frailty model, subjects in the same group share the same frailty value which generates dependence between those subjects who share frailties.

The shared frailty model can be written as follows:

$$\rho_{ik}(t|u_i) = u_i \rho_{ik}(t) = \rho_0(t) \exp(\mathbf{z}'_{ik}\boldsymbol{\beta} + u_i),$$

where $\rho_{ik}(t)$ is the conditional hazard rate function for the kth subject from the ith cluster conditional on u_i, $\rho_0(t)$ is the baseline hazard, $\boldsymbol{\beta}$ is the fixed effects vector of dimension p, $\mathbf{z}_{ik}$ is the vector of covariates, and u_i is the random effect for the ith cluster. Thus, subjects in the same cluster i share the same frailty factor and it is a conditional hazard model, given the u_i. The cluster may represent a family or a single subject for which multiple episodes are observed.

The distribution of u_i may be Gamma, Gaussian, or another distribution. The gamma distribution is often chosen because of its mathematical tractability and because it is widely used. The one-parameter gamma distribution is defined as:

$$f_w(u) = \frac{v^{1/\theta - 1} e^{(u/\theta)}}{\theta^{1/\theta} \Gamma(1/\theta)}$$

with Γ the gamma function and $\mathrm{E}(u) = 1$ and $\mathrm{Var}(u) = \theta$. This means that subjects in class i with $u_i > 1$ are frail (having a higher risk) while subject with $u_i < 1$ are strong (having a lower risk). The parameter θ gives information on the clusters or classes heterogeneity in the population.

5.3 Illustrations

5.3.1 Poisson Regression Data (N = 1000)

Poisson regression is a commonly used count response regression model where events are considered to be of the same kind. Since the model is the '*foundation*' of other recurrent event models, we will look at some of the models' behavior. Few real-life datasets are truly Poisson, where the mean and variance are equal. The vast majority of datasets that are initially thought to be close to Poisson usually have a larger variance than the mean. When this is the case, we say that the Poisson model is overdispersed, which may cause the standard errors of the estimates to be underestimated. When this is the case, a variable may appear to be a significant predictor when in fact it is not. We will illustrate these behaviors by generating data that follows a Poisson regression model, then remove a predictor and see the effect this has on the Poisson model.

We are going to generate $\mathrm{n} = 1000$ standard normally distributed observation for each of 3 independent variables Z_1, Z_2, and Z_3, and then apply the linear equation $\mathbf{Z}'\boldsymbol{\beta} = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3$ with coefficient $(\beta_1, \beta_2, \beta_3) = (-0.50, -0.50, -0.25)$. After exponentiating $\mathbf{Z}'\boldsymbol{\beta}$ the Poisson variate y is generated using the probability

Table 5.1 General linear Poisson model on y with independent variables Z_1, Z_2, and Z_3

```
Generalized linear models                         No. of obs      =      1,000
Optimization     : ML                             Residual df     =        996
                                                  Scale parameter =          1
Deviance         =  1129.256566                   (1/df) Deviance =   1.133792
Pearson          =  1085.205183                   (1/df) Pearson  =   1.089563
Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = Ln(u)                   [Log]
                                                  AIC             =   2.629619
Log likelihood   = -1310.809415                   BIC             =  -5750.868
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          z1 |  -.4896427   .0270418   -18.11   0.000    -.5426436   -.4366417
          z2 |  -.4291967   .0286559   -14.98   0.000    -.4853613   -.3730322
          z3 |  -.2327014   .0274627    -8.47   0.000    -.2865273   -.1788755
       _cons |  -.0219569    .034287    -0.64   0.522    -.0891583    .0452444
------------------------------------------------------------------------------
```

Table 5.2 General linear Poisson model on y with independent variables Z_2 and Z_3

```
Generalized linear models                         No. of obs      =      1,000
Optimization     : ML                             Residual df     =        997
                                                  Scale parameter =          1
Deviance         =  1451.594534                   (1/df) Deviance =   1.455962
Pearson          =  1503.530661                   (1/df) Pearson  =   1.508055
Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = Ln(u)                   [Log]
                                                  AIC             =   2.949957
Log likelihood   = -1471.978399                   BIC             =  -5435.437
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          z2 |  -.4736272   .0284366   -16.66   0.000    -.5293618   -.4178925
          z3 |  -.2235353   .0276194    -8.09   0.000    -.2776683   -.1694023
       _cons |   .1155962   .0311098     3.72   0.000     .0546222    .1765703
------------------------------------------------------------------------------
```

integral transform methods of Kemp and Kemp (1990, 1991) and the method of Kachitvichyanukul (1982) through the Stata software.

The Poisson variate y is next modeled on the three randomly generated independent variables Z_1, Z_2, and Z_3. The results of the analysis are presented in Table 5.1.

Although a sample size of n = 1000 usually is considered to be a decent sample size for a clinical study with one single treatment arm, we will find some interesting outcomes from the analysis results. We find that the estimates of the coefficients (β_1, β_2, β_3) are (−0.4896, −0.4292, −0.2327). All parameter estimates are lower than what we assigned them to be, especially the estimate of β_2 which is −0.4292 instead of −0.50. The Pearson dispersion statistic, defined as the Pearson statistic divided by the model degrees of freedom, would be equal to 1.0 if the model is the '*correct*' one. Here the Pearson statistic is 1.089563, which is about 9% higher than expected.

We will now omit predictor Z_1 and again model the data on the remaining variables. Thus, the Poisson variate y is then modeled on the two randomly generated independent variables Z_2 and Z_3. The results of the analysis are presented in Table 5.2.

We find that the estimates of the coefficients (β_2, β_3) are (−0.4736, −0.2235). Both parameter estimates are still lower than what we assigned them to be, but not worse than from the previous model. The Pearson dispersion statistic has

now increased from 1.089563 to 1.508055. Here the Pearson statistic has notably increased, telling us that the model is overdispersed. The AIC and BIC statistics are also inflated.

5.3.2 *Poisson Regression Data (N = 25,000)*

We will now increase the sample size to 25,000, but everything else will be kept the same. When we use all of the independent variables Z_1, Z_2, and Z_3 in the model (table not shown), we find what we can expect, that the Pearson dispersion statistic is 1.0121, which is very close to 1.0. The estimates of the coefficients (β_1, β_2, β_3) are (-0.4889, -0.5058, -0.2563), i.e., the parameter estimates are very close to what we assigned them to be.

Again, we now omit predictor Z_1 and again model the data on the remaining variables. Thus, the Poisson variate y is then modeled on the two randomly generated independent variables Z_2 and Z_3. The results of the analysis are presented in Table 5.3.

We find that the estimates of the coefficients (β_2, β_3) are (-0.5062, -0.2565). Both parameter estimates are still close to what we assigned them to be, though this is not an indication that the model is appropriate. What has notably changed is that the Pearson dispersion statistic is now about 1.375, telling us that the model is overdispersed. Given the very large dataset of 25,000 observations, the dispersion statistic correctly indicates that the Poisson model is overdispersed. We see that the model obviously does not fit the data.

We will now assume that the variance is proportional rather than equal to the mean, and estimate the scale parameter φ dividing Pearson's chi-squared by its degrees of freedom (df), which gives us the value 1.375. We see that the variance is about 37.5% larger than the mean. This means that we should adjust the standard errors multiplying by 1.173, the square root of 1.375 (see Table 5.4).

Using this procedure of scaling the standard errors we have essentially attributed all the lack of fit to pure error. We can also try to run the Poisson model with the

Table 5.3 General linear Poisson model on y with independent variables Z_2 and Z_3

```
Generalized linear models                         No. of obs      =     25,000
Optimization     : ML                             Residual df     =     24,997
                                                  Scale parameter =          1
Deviance         =  34918.46873                   (1/df) Deviance =   1.396906
Pearson          =  34373.65875                   (1/df) Pearson  =   1.375111
Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = Ln(u)                   [Log]
                                                  AIC             =   2.953739
Log likelihood   =  -36918.7331                   BIC             =  -218216.9
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          z2 |  -.5062168    .005487   -92.26   0.000    -.5169712   -.4954625
          z3 |  -.2564794   .0054965   -46.66   0.000    -.2672522   -.2457065
       _cons |   .1130332   .0063441    17.82   0.000     .1005989    .1254674
------------------------------------------------------------------------------
```

Table 5.4 General linear Poisson model on y with independent variables Z_2 and Z_3

```
Generalized linear models                         No. of obs      =     25,000
Optimization     : ML                             Residual df     =     24,997
                                                  Scale parameter =          1
Deviance         =  34918.46873                   (1/df) Deviance =   1.396906
Pearson          =  34373.65875                   (1/df) Pearson  =   1.375111
Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = Ln(u)                   [Log]
                                                  AIC             =   2.953739
Log likelihood   =  -36918.7331                   BIC             =  -218216.9
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          z2 |  -.5062168   .0064344   -78.67   0.000     -.518828   -.4936057
          z3 |  -.2564794   .0064454   -39.79   0.000    -.2691122   -.2438466
       _cons |   .1130332   .0074394    15.19   0.000     .0984522    .1276142
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion.)
```

robust option to compute standard errors using the robust or '*sandwich*' estimator. Doing so we will get very similar results. In either case, all tests have to be done using Wald's statistic. Likelihood ratio tests are not possible because we are not making full distributional assumptions about the outcome, relying instead on assumptions about the mean and variance.

5.3.3 *Negative Binomial Regression (N = 25,000)*

Since the Poisson model with the two independent variables Z_2 and Z_3 was overdispersed, we will now fit a negative binomial model to the recent data with the same variables Z_2 and Z_3. The results are shown in Table 5.5.

The alpha in Table 5.5 is the variance of the multiplicative random effect. We have overwhelming evidence of overdispersion. For testing hypotheses about the regression coefficients, we can use either Wald tests or likelihood ratio tests.

Table 5.5 Negative Binomial model on y with independent variables Z_2 and Z_3

```
Negative binomial regression                      Number of obs     =     25,000
                                                  LR chi2(2)        =    6437.40
Dispersion     = mean                             Prob > chi2       =     0.0000
Log likelihood = -36210.117                       Pseudo R2         =     0.0816
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          z2 |  -.5068882   .0067312   -75.30   0.000    -.5200811   -.4936952
          z3 |    -.25639   .0066132   -38.77   0.000    -.2693517   -.2434284
       _cons |   .1127922   .0071559    15.76   0.000      .098767    .1268175
-------------+----------------------------------------------------------------
    /lnalpha |  -1.342182   .0367011                     -1.414114   -1.270249
-------------+----------------------------------------------------------------
       alpha |    .261275   .0095891                      .2431408    .2807617
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 1417.23             Prob >= chibar2 = 0.000
```

Table 5.6 Comparing estimates and standard errors side by side

Variable	poisson Table 3.2.1.	overdisp Table 3.2.2.	nbreg Table 3.3.1.
y			
z2	-.50621684	-.50621684	-.50688815
	.00548703	.00643438	.0067312
z3	-.25647936	-.25647936	-.25639002
	.00549646	.00644543	.00661321
_cons	.1130332	.1130332	.11279225
	.00634412	.00743944	.00715588
lnalpha			
_cons			-1.3421817
			.03670105

5.3.4 Comparing Estimates and Standard Errors

The parameter estimates based on the negative binomial model are not very different from those based on the Poisson regression model. We will now compare the models side by side in Table 5.6.

Both sets of parameter estimates would lead to the same conclusions. Looking at the standard errors reported just below the coefficient estimates, we see that both approaches to overdispersion lead to very similar estimates and that ordinary Poisson regression underestimates the standard errors.

5.3.5 Goodness of Fit

We will evaluate the goodness of fit using the second dataset above with 25,000 observations. One way to compute the deviance of the negative binomial model is to feed the estimate of the variance into the generalized linear model. The deviance statistic is now 1.0741, which tells us that the negative binomial model fits much better than the Poisson model, but still, has a deviance just above the five percent value. One way to model this type of situation is to assume that the data come from a mixture of two populations, one where the counts are always zero, and another population where the count has a Poisson distribution with mean μ. In this, model zero counts can come from either population, while positive counts come only from the second population.

The distribution of the outcome can then be modeled in terms of two parameters, π the probability of '*always zero*', and μ, the mean number of for those not in the '*always zero*' population. A natural way to introduce covariates is to model the logit of the probability π of always zero and the log of the mean μ for those not in the always zero population.

5.3.6 Simulations

Clinical trial simulation studies can be used to assess the impact of many aspects of trial design, conduct, analysis and decision making. Simulation studies can play a vital role in improving the efficiency of drug development within the pharmaceutical industry, but only if they are well designed and conducted. An efficient way of evaluating the properties that different models have for the study design and analysis that we are considering is to use simulations. A number of common software packages make this possible, such as EAST, SAS, Stata, and R.

A comprehensive overview is given of how to use simulations for designing clinical trials and how to analyze the simulated clinical trial data in Ette et al. (2002). A generic template for clinical trials simulations that are typically required by statisticians has been developed by Westfall et al. (2008). Realistic clinical trials datasets are created using a unifying model that allows general correlation structures for endpoint and timepoint data and nonnormal distributions (including time-to-event), and computationally efficient algorithms are presented. The structure allows for patient dropout and noncompliance. A grid-enabled SAS-based system has been developed to implement this model and details are presented summarizing the system development (Westfall et al. 2008, 2010).

For instance, we may use simulations to compare the conditional frailty model and several variance-corrected and frailty models with a known data generating process that exhibits heterogeneity, event dependence, both, and neither. Box-Steffensmeier and De Boef (2006) did this and focused their simulations on the comparison of the three more popular and promising variance-corrected models: the Andersen–Gill, conditional gap time, and conditional elapsed time models, and the basic frailty model estimated with a gamma random effect. They gauged model performance on three dimensions: the bias in the estimated treatment effects as well as in the estimated variance of the random effect, bias in the standard errors, and rate of which the estimated standard errors includes the true parameter. Their simulations suggested that the conditional frailty model can estimate the effects of both sources of correlation simultaneously and retrieve the parameters of the true data generating process better in all four cases. Furthermore, in the simulations they investigated, the conditional frailty model performed similarly to, or better than, the variance-corrected and frailty alternatives. In the case of both heterogeneity and event dependence, only the conditional frailty model performed well. So, in cases where there is a possibility of both, and often we cannot rule either out, the conditional frailty model is recommended.

5.4 Discussion

The foundation of recurrent event analysis is survival analysis has been a common and well-accepted strategy to study treatment effect in a population of patients. During

the last few years, there has been an increasing interest in assessing therapy effect not only by using time to death, but also time to surrogate events such as time to hospitalization. The combined endpoint of time to death and time to disease-related hospitalizations is often analyzed with a time-to-first-event analysis, which has the drawback of waste of information and indistinct handling of two clinically different events.

The analysis of multiple events per subject cannot be approached by a standard Cox model, where the assumption of independence of observations is not valid. In order to account for intra-subject correlation, we have presented the use of marginal and multistate models using a counting process approach for, for instance, the joint analysis of survival and time to disease-related hospitalizations.

In a comparison of common statistical methods for analyzing recurrent event data, the results with each method for lack of bias, efficiency, and robustness for within-subject correlation are not, but depending on the process driving the event counts. In general, the Poisson regression with correction for overdispersion has similar coverage probabilities of confidence intervals, but slightly higher type I error rates compared to the robust Andersen–Gill and negative binomial approaches, which are therefore preferable. Advantages in power for some situations are only at the price of an increased type I error. The negative binomial regression surprisingly produces results similar to those of the Andersen–Gill approach, even when the distribution is not homogeneously Poisson. On the other hand, for homogeneous Poisson processes, the Andersen–Gill approach does not lose efficiency in comparison with the perfectly fitting negative binomial regression model (Jahn-Eimermacher et al. 2015). The demonstrated comparability of the Andersen–Gill approach and negative binomial regression for Poisson processes supports the findings of Metcalfe and Thompson (2006). The results are in agreement with the data example presented by Guo et al. (2008), in which trial results from an Andersen–Gill model were similar to those from Poisson regression.

For the conditional model not derived from the Poisson process, with all the investigated methods, estimation of a zero treatment effect and its standard error may be considered as acceptable, and thus be applicable to hypothesis testing. However, the effect estimates are biased, whatever method is used. All of the investigated methods are not applicable if the independent increment assumption is violated. For a specific application, this assumption, therefore, must be checked by appropriate sensitivity analyses. So, results could be compared with those of the conditional model of Prentice et al. (1981) or the marginal model of Wei et al. (1989). However, these approaches also have sources of bias as demonstrated by Therneau and Grambsch (2000) and Kelly and Lim (2000) and, furthermore, the applicability of the marginal model to recurrent failure time data is discussed critically in Metcalfe and Thompson (2007).

We found no advantages in performance with Poisson regression as compared with the Andersen–Gill approach, which allows more complex analyses and may, therefore, be preferable. The Poisson regression remains applicable when only aggregated event counts are available or when the actual time of occurrence of an event cannot be determined. Dean and Balshaw (1997) demonstrated for nonhomogeneous Pois-

son processes that treatment effects can be efficiently estimated based on aggregated count data as long as censoring is balanced between treatment groups.

Standard errors might be substantially underestimated with all the methods examined if within-subject correlation is not accounted for, in accordance with previous findings (Glynn and Buring 1996, Therneau and Hamilton 1997, Metcalfe and Thompson 2006). Robust variance estimation can be used to adjust for the simulated degree of within-subject-correlation, however, in rare cases, data may be even more highly correlated (Thall 1988). In those situations, the robust methods may also fail to prevent type I error from increasing to unacceptable levels.

The use of a gamma distribution for the random effect is common in the literature (Stukel 1993; Metcalfe and Thompson 2006; Thomsen and Parner 2006). Regression parameter estimation in a gamma frailty model seems to be robust to frailty distribution misspecification as Hsu et al. (2007) demonstrated for single event data in cohort and case-control family trials. Kelly and Lim (2000), Therneau and Grambsch (2000) and Metcalfe and Thompson (2006) used realizations from normal and uniform distributions, with which the Andersen–Gill method underestimated treatment effects.

Finally, the most appropriate model should be chosen based on the anticipated nature and structure of the data.

References

Aalen, O. O., Borgan, Ø. & Gjessing, H. K. (2008). *Survival and event history analysis: A process point of view*. Springer.

Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine, 19,* 453–473.

Andersen, P. K., Abildstrom, S. Z., & Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research, 11,* 203–215.

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer-Verlag.

Andersen, P. K., Borgan, Ø., Gill, R. D., & Keiding, N. (1997). *Statistical models based on counting processes* (Corrected ed.). New York: Springer.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics, 10,* 1100–1120.

Andersen P. K., & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, *11*, 91.

Blossfeld, H., & Rohwer, G. (1995). *Techniques of event history modeling*. New Jersey: Lawrence Erlbaum.

Box-Steffensmeier, J. M., & De Boef, S. (2006). Repeated events survival models: The conditional Frailty model. *Statistics in Medicine, 25,* 3518–3533.

Brenner, H., & Gefeller, O. (1997). Deriving more up-to-date estimates of long-term patient survival. *Journal of Clinical Epidemiology, 50,* 211–216.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review/Revue Internationale de Statistique*. *43*(1), 45–57.

Coleman, M. P., Forman, D., Bryant, H., Butler, J., Rachet, B., Maringe, C., Nur, U., et al. (1995–2007). Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK,

(the International Cancer Benchmarking Partnership): An analysis of population-based cancer registry data. *Lancet, 377*: 127–138.
Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis, 5,* 315–327.
Cook, R. J., & Lawless, J. F. (2002). Analysis of repeated events. *Statistical Methods in Medical Research, 11,* 141–166.
Cook, R. J., & Lawless, J. F. (2007). *The statistical analysis of recurrent events*. Springer.
Courgeau, D., & Lelièvre, E. (1992). *Event history analysis in demography*. Oxford: Clarendon.
Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B, 34*(2), 187–220.
Crowther, M. J. (2016). *MULTISTATE: Stata module to perform multi-state survival analysis*. https://ideas.repec.org/c/boc/bocode/s458207.html.
Dean, C. B., & Balshaw, R. (1997). Efficiency lost by analyzing counts rather than event times in Poisson and overdispersed Poisson regression models. *Journal of the American Statistical Association, 92,* 1387–1398.
Duchateau, L., & Janssen, P. (2008). *The Frailty model*. Springer.
Efron, B. (1974). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association., 72*(359), 557–565.
Ette, E. I., Godfrey, C. J., Ogenstad, S., & Williams, P. (2002). Analysis of simulated clinical trials. In *Simulation for designing clinical trials: A pharmacokinetic-pharmacodynamic modeling perspective.* Marcel-Dekker, IBSN: 0-8247-0862-8.
Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association, 94,* 496–509.
Fuchs, H. J., Borowitz, D. S., Christiansen, D. H., Morris, E. M., Nash, M. L., Ramsey, B. W., Rosenstein, B. J. et al. (1994). Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. The Pulmozyme Study Group. New England Journal of Medicine, *331*(10):637–42.
Gamel, J. W., & Vogel, R. L. (2001). Non-parametric comparison of relative versus cause-specific survival in Surveillance, Epidemiology and End Results (SEER) programme breast cancer patients. *Statistical Methods in Medical Research, 10,* 339–352.
Ghosh, D., & Lin, D. Y. (2000). Nonparametric analysis of recurrent events and death. *Biometrics, 56*(2), 554–562.
Glynn, R. J., & Buring, J. E. (1996). Ways of measuring rates of recurrent events. *British Medical Journal, 312,* 364–367.
Guo, Z., Gill, T. M., & Allore, H. G. (2008). Modeling repeated time-to-event health conditions with discontinuous risk intervals. *Methods of Information in Medicine, 47,* 107–116.
Gutierrez, R. G. (2002). Parametric frailty and shared frailty survival models. *Stata Journal, 2,* 22–44.
Hilbe, J. M. (2007). *Negative binomial regression*. Cambridge University Press.
Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika, 71,* 75–83.
Hougaard, P. (1986a). A class of multivariate failure time distributions. *Biometrika, 73,* 671–678.
Hougaard, P. (1986b). Survival models for heterogeneous populations derived from stable distributions. *Biometrika, 73*, 397–96.
Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis, 1,* 255–273.
Hougaard, P. (1999). Multi-state models: A review. *Lifetime Data Analysis, 5,* 239–264.
Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer.
Hsu, L., Gorfine, M., & Malone, K. (2007). On robustness of marginal regression coefficient estimates and hazard functions in multivariate survival analysis of family data when the frailty distribution is mis-specified. *Statistics in Medicine, 26,* 4657–4678.
Jahn-Eimermacher, A., Ingel, K., Ozga, A. K., Preussler, S., & Binder, H. (2015). Simulating recurrent event data with hazard functions defined on a total time scale. *BMC Medical Research Methodology, 15,* 16.

Kachitvichyanukul, V. (1982). *Computer generation of poisson, binomial, and hypergeometric random variables*. Ph.D. thesis, Purdue University.

Kelly, P. J., & Lim, L. L. (2000). Survival analysis for recurrent event data: An application to child-hood infectious diseases. *Statistics in Medicine, 19,* 13–33.

Kemp, A. W., & Kemp, C. D. (1990). A composition-search algorithm for low-parameter Poisson generation. *Journal of Statistical Computation and Simulation, 35,* 239–244.

Kemp, C. D., & Kemp, A. W. (1991). Poisson random variate generation. *Applied Statistics, 40,* 143–158.

Lambert, P. C., & Royston, P. (2009). Further development of flexible parametric models for survival analysis. *Stata Journal, 9,* 265–290.

Li, Q., & Lagakos, S. (1997). Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and terminating event. *Statistics in Medicine, 16,* 925–940.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine, 13,* 2233–2247.

Lin, D. Y., & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association, 84,* 1074–1078.

Lunn, M., & McNeil, D. (1995). Applying Cox regression to competing risks. *Biometrics, 51,* 524–532.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Kristensen, P., Mann, J. F. E., et al. (2016). Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine, 375,* 311–322.

Metcalfe, C., & Thompson, S. G. (2006). The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Statistics in Medicine, 25,* 165–179.

Metcalfe, C., & Thompson, S. G. (2007). Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: Should it be applied to recurrent events outcome? *Statistical Methods in Medical Research, 16,* 103–122.

Møller, H., Sandin, F., Bray, F., Klint, A., Linklater, K. M., Purushotham, A., et al. (2010). Breast cancer survival in England, Norway and Sweden: A population-based comparison. *International Journal of Cancer, 127,* 2630–2638.

Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika, 68,* 373–379.

Royston, P., & Lambert, P.C. (2011). *Flexible parametric survival analysis using Stata: Beyond the Cox model*. Stata-Press.

Royston, P., & Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine, 21,* 2175–2197.

Sagara, I., Giorgi, R., Doumbo, O. K., Piarroux, R., & Gaudart, J. (2014). Modelling recurrent events: Comparison of statistical models with continuous and discontinuous risk intervals on recurrent malaria episodes data. *Malaria Journal, 13,* 293.

Stukel, T. A. (1993). Comparison of methods for the analysis of longitudinal interval count data. *Statistics in Medicine, 12,* 1339–1351.

Thall, P. F. (1988). Mixed Poisson likelihood regression models for longitudinal interval count data. *Biometrics, 44,* 197–209.

Therneau, T. M. (1997). Extending the Cox model. In *Proceedings of the First Seattle Symposium in Biostatistics*. New York: Springer.

Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data*. New York: Springer.

Therneau, T. M., & Hamilton, S. A. (1997). rhDNase as an example of recurrent event analysis. *Statistics in Medicine Banner, 16,* 2029–2047.

Thomsen, J. L., & Parner, E.T. (2006). Methods for analysing recurrent events in health care data. Examples from admissions in Ebeltoft Health Promotion project. *Family Practice, 23*, 407–413.

Tsiatis, A. A. (2005). Competing risks. In P. Armitage & T. Colton (Eds.), *Encyclopedia of Biostatistics* (2nd ed., pp. 1025–1035). Hoboken, NJ: Wiley.

Vaupel, J. W., Manton, K., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography, 16,* 439–454.

Wei, L. J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association., 84,* 1065–1073.

Westfall, P. H., Dmitrienko, A., DeSouza, C., & Tomoiaga, A. (2010). Clinical trials simulation: A publicly available, grid-enabled, GUI-Driven SAS® system. SAS Global Forum 2010, Paper 180–2010.

Westfall, P. H., Tsai, K., Ogenstad, S., Tomoiaga, A., Moseley, S., & Lu, Y. (2008). Clinical trials simulation: A statistical approach. *Journal of Biopharmaceutical Statistics, 18*(4), 611–630.

Wienke, A., Holm, H., Skytthe, A., & Yashin, A. (2001). The heritability of mortality due to heart diseases: A correlated Frailty model applied to Danish twins. *Twin Research, 4*(4), 266–274.

Chapter 6
Response-Adaptive Allocation for Binary Outcomes: Bayesian Methods from the BASS Conference

Roy T. Sabo

6.1 Introduction

Outcome- or response-adaptive allocation methods are used to adjust randomization probabilities in clinical trials based on observations from previously accrued patients. These methods aim to achieve one of several allocation goals, which have included maximizing statistical power, balancing for covariates, and maximizing treatment benefit. In the latter case, adaptive allocation strategies aim to treat patients as ethically as possible, often by minimizing the expected number of treatment failures. These "optimal designs" achieve this minimization through algorithms and functions of success probabilities in each group of subjects.

Though much of the conceptual and theoretical work in adaptive allocation methods has been conducted in the frequentist framework, Bayesian methods are a natural fit for conducting outcome-adaptive allocation in practice. These methods are more easily adaptable to small sample cases and are generally more flexible than are frequentist alternatives. For instance, frequentist allocation approaches generally require an initial *lead-in* period where allocation probabilities are held constant in order to overcome small-sample irregularities in proportion estimates. Some researchers have introduced scaling parameters into allocation algorithms that restrict allocation in early phases of a trial and gradually allow increasing adaptation, but even these approaches cannot account for situations where a treatment group has no observed successes, which would result in no allocation to that group. Bayesian methods can overcome these difficulties in several ways, most notably through informative prior specification or through replacing success proportion estimates with posterior or predictive probabilities of treatment superiority. Bayesian methods are also more readily adapted to account for situations where allocation ratios are desired to adapt

R. T. Sabo (✉)
Department of Biostatistics, Virginia Commonwealth University,
830 East Main Street, Richmond, VA 23298-0032, USA
e-mail: roy.sabo@vcuhealth.org

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_6

based on information from multiple outcomes, as joint distributions between multiple outcomes can be estimated through a posterior distribution in a straightforward manner.

In this Chapter we provide two examples of Bayesian approaches to outcome-adaptive allocation. The first overcomes the necessity of a *lead-in* by eliciting an informative yet skeptical prior that exhibits decreasing influence on the posterior as more patients enter a trial. This approach – dubbed the *Decreasingly Informative Prior* approach – was the subject of a 2013 presentation at the Biopharmaceutical and Applied Statistics Symposium (BASS) as well as a subsequent publication Sabo (2014). The second method presents an approach to base allocation upon two outcomes simultaneously, such as in trials where both treatment efficacy and toxicity are important. This approach was the subject of a 2012 BASS presentation and subsequent publication Sabo et al. (2013). In both cases we focus on two- and three-group clinical trials with binary outcomes. A general review of response-adaptive allocation will be provided in the next section, while the Bayesian approach will be covered in Sect. 6.3. The decreasingly informative prior approach will be discussed in Sect. 6.4, while the two-outcome approach will be presented in Sect. 6.5.

6.2 Response-Adaptive Allocation

6.2.1 Optimal Allocation

Rosenberger et al. (2001) derived optimal allocation weights for two-group trials with binary outcomes, with the goal to minimize the expected number of treatment failures. These weights are given below in Eq. 6.1.

$$\begin{aligned} w_1 &= \frac{\sqrt{p_1}}{\sqrt{p_1} + \sqrt{p_2}}, \\ w_2 &= 1 - w_1, \end{aligned} \tag{6.1}$$

where p_j is the proportion of successfully treated patients in group j $(j = 1, 2)$, and where weight w_j is the probability the next patient will be allocated into the jth treatment group. In practice the unknown estimates p_1 and p_2 are replaced with the current sample proportions $\hat{p}_1$ and $\hat{p}_2$, which could lead to the awkward scenario in early phases of a trial where the weights given in Eq. 6.1 are incalculable due to no events being observed in either of the two groups.

Optimal allocation ratios for three-group trials were established numerically by Tymofyeyev et al. (2007) and in closed-form by Jeon and Hu (2010). These optimal allocation ratios depend upon the relative magnitudes of the success proportions in each group and a constant $B \in (0, 1/3)$, which is a lower allocation bound selected by the investigator (Jeon and Hu recommend selecting $0 < B \leq 1/3$ to prevent situations where a treatment ends up with no patients). We present them here with

minor corrections due to typos in the original manuscript. Let p_1, p_2 and p_3 be the true efficacy rates of treatments 1, 2 and 3, and let $\mathbf{w}^* = (w_1^*, w_2^*, w_3^*)^T$ denote the vector of optimal allocation proportions. Then for $p_1 > p_2 > p_3$, $B \in (0, 1/3)$, and $q_j = 1 - p_j$, $j = 1, 2, 3$, the allocation rates are

$$\begin{aligned} w_1^* &= l_2^{-1}(l_1 + l_3 B) \\ w_2^* &= B \\ w_3^* &= 1 - B - w_1^*, \end{aligned} \tag{6.2}$$

where,

$$\begin{aligned} l_1 &= \frac{a(p_1 - p_3) + b(p_2 - p_3) + d}{p_3 q_3}, \\ l_2 &= \frac{b(p_1 - p_2) + c(p_1 - p_3) - d}{p_1 q_1} + l_1, \\ l_3 &= \frac{a(p_1 - p_2) - c(p_2 - p_3) + d}{p_2 q_2} - l_1, \\ a &= -\frac{Bq_2 - (B-1)q_3}{p_1 q_1} \\ b &= -\frac{B(q_3 - q_1)}{p_2 q_2} \\ c &= \frac{Bq_2 - (B-1)q_1}{p_3 q_3} \\ d &= \sqrt{-ab(p_1 - p_2)^2 - ac(p_1 - p_3)^2 - bc(p_2 - p_3)^2}. \end{aligned}$$

If $w_1^* > B$ and $w_3^* > B$ then (Eq. 6.2) is the optimal solution. If $w_1^* \leq B$, the solution is $\mathbf{w}^* = (B, B, 1 - 2B)^T$. If $w_3^* \leq B$, the solution is $\mathbf{w}^* = (1 - 2B, B, B)^T$. When $p_1 = p_2 > p_3$ the solution is:

$$w_1^* = w_2^* = \frac{\sqrt{p_1}}{2(\sqrt{p_1} + \sqrt{p_3})}, \; w_3^* = \frac{\sqrt{p_3}}{\sqrt{p_1} + \sqrt{p_3}},$$

provided $w_j^* \geq B \, \forall \, j$. If $B > \frac{\sqrt{p_1}}{2(\sqrt{p_1}+\sqrt{p_3})}$, the solution is $\mathbf{w}^* = (B, B, 1 - 2B)^T$. If $B > \frac{\sqrt{p_3}}{\sqrt{p_1}+\sqrt{p_3}}$, the solution is $\mathbf{w}^* = ((1 - B)/2, (1 - B)/2, B)^T$. When $p_1 > p_2 = p_3$ the solution is:

$$w_1^* = \frac{\sqrt{p_1}}{\sqrt{p_1} + \sqrt{p_3}}, \; w_2^* = w_3^* = \frac{\sqrt{p_3}}{2(\sqrt{p_1} + \sqrt{p_3})},$$

provided $w_j^* \geq B \, \forall \, j$. If $B > \frac{\sqrt{p_3}}{2(\sqrt{p_1}+\sqrt{p_3})}$, the solution is $\mathbf{w}^* = (1 - 2B, B, B)^T$.

6.2.2 Natural Lead-In

Thall and Wathen (2007) model the root for the two-group case as an increasing function of the observed sample size $(n/2N)$, where n is the number of observed patients and N is the planned total sample size. Here, the weighting algorithm becomes

$$\begin{aligned} w_1 &= \frac{p_1^{n/2N}}{p_1^{n/2N} + p_2^{n/2N}}, \\ w_2 &= 1 - w_1. \end{aligned} \tag{6.3}$$

This approach has the effect of acting as a *natural* lead-in, since it forces equal weights at the beginning of a trial and gradually allows more adaptation as the trial continues. In addition, as $n \to N$ the weights in Eq. 6.3 approach the same structural form as those given in Eq. 6.1.

In the three-group case Hu and Zhang (2004) introduced an allocation function based on the doubly adaptive biased coin design (Eisele 1994), which is given as follows

$$\begin{aligned} w_j &= \frac{w_j^* \left((w_j^* \sum_{i=1}^3 n_i)/n_j \right)^\gamma}{\sum_{k=1}^3 w_k^* \left((w_k^* \sum_{i=1}^3 n_i)/n_k \right)^\gamma} \\ j &= 1, 2, 3, \end{aligned} \tag{6.4}$$

where n_j is the current observed sample size in group j, w_j^* is the current optimal allocation weight (Eq. 6.2) in group j, and γ is a tuning parameter for calibrating the degree of randomness of the allocation probability function. By setting $\gamma = (N - (n+1))/n$ we again achieve a *natural* lead-in that forces equal allocation early in the trial, and approaches the optimum allocation rates found in Eq. 6.2 as $n \to N - 1$ (Bello and Sabo 2016).

6.3 General Bayesian Approach

In the Bayesian framework proportions p_j for treatment groups $j = 1, \ldots, k$, are assigned a common prior distribution $\pi(\theta_0)$, where $\pi(.)$ is some distributional form and θ_0 is some fixed value. The prior distributions are combined with likelihood distributions $p(y_j|p_j, n_j)$ for each treatment group, where $p(.)$ is some distributional form, n_j is the number of observed patients in treatment group j, and y_j is the number of "successful" events observed in n_j subjects. The specific choice of prior and likelihood are then synthesized into a posterior distribution for parameter p_j

$$P(p_j|y_j, n_j, \theta_0) \propto p(y_j|p_j, n_j)\pi(\theta_0),\ j = 1, \ldots, k. \tag{6.5}$$

We can slightly generalize this framework by establishing a hierarchical posterior distribution for any parameter θ as follows

$$\theta \sim P(\theta|y) = \frac{p(y|\theta, n)\pi(\theta|\theta_0, n, N)g(\theta_0|\lambda)}{\int p(y|\theta, n)\pi(\theta|\theta_0, n, N)g(\theta_0|\lambda)}, \tag{6.6}$$

where y are the observed data, $p(.)$ is the likelihood function, $\pi(.|\theta_0, n, N)$ is the prior information on θ, and $g(.)$ is a hyperprior on θ_0 with hyperparameter λ. This posterior can be used to estimate the mean or mode success rate in each group, which can then be used in Eqs. 6.1 or 6.2.

6.3.1 Posterior Estimates and Probabilities

As an alternative to posterior means or modes, Huang et al. (2007) and Thall and Wathen (2007) replaced success probabilities with probabilities of greater treatment response. Here we calculate the posterior probability that p_1 is greater than p_2, so that allocation weights increase in favor of treatment 1 as evidence of its superiority accumulates. While similar to using success rates directly, these probabilities tend to provide quicker and greater adaptation. In two-arm trials (Thompson 1933; Thall and Wathen 2007) we need only calculate one probability

$$\begin{aligned} P_1 &= P(p_1 > p_2|y, n, \theta_0) \\ P_2 &= 1 - P_1, \end{aligned} \tag{6.7}$$

where $y = (y_1, y_2)$ and $n = (n_1, n_2)$. In three-arm trials (Bello and Sabo 2016; Sabo and Bello 2017) we calculate three probabilities

$$\begin{aligned} P_1 &= [(p_1 > p_2) \cap (p_1 > p_3)|y, n, \theta_0], \\ P_2 &= [(p_2 > p_1) \cap (p_2 > p_3)|y, n, \theta_0], \\ P_3 &= [(p_3 > p_1) \cap (p_3 > p_2)|y, n, \theta_0] \end{aligned} \tag{6.8}$$

where $y = (y_1, y_2, y_3)$ and $n = (n_1, n_2, n_3)$. In practice, these posterior probabilities can be used in place of the unknown population success rate for the corresponding group.

Predictive probabilities could also be used in adaptive allocation (Sabo and Bello 2017). Many predictive probability approaches in clinical trials use the current posterior probability distribution (as given in Eq. 6.6) as the new prior, and combine this information with some likelihood for the patients who have yet to accrue or whose outcomes are currently unobserved, and the resulting predictive distributions are used to calculate the probability of interest. Using the standard formulation of the

predictive distribution produces similar mean or mode estimates to those obtained from simulating from the posterior distribution, since both the posterior and predictive posterior distributions have the same center. An alternative approach, as outlined in Sabo and Bello (2017), relies upon the *re-use* of skeptical prior information to calculate predictive probabilities. Rather than assume that future patients will behave similarly to patients already accrued into the trial, we return to our skeptical assumptions expressed in the prior distribution $\pi(\theta_0)$ to conservatively account for uncertainty in the non-accrued patients. The rationale for using this skeptically predictive approach is to avoid the assumption that there are no time-based biases in patient accrual or treatment effectiveness, an issue raised by Korn and Freidlin (2011) in their critique of outcome-adaptive allocation. In essence, the predictive probability distribution is used to simulate responses y_j^* for the remaining n_j^* subjects in treatment j. Direct sampling or markov-chain monte carlo methods (with T iterations) can be used to estimate predictive probabilities for between-treatment comparisons as

$$\begin{aligned} P_1 &= P(p_1 > p_2|\theta_0, y, y^*, n, n^*) = \sum_{t=1}^{T} I(p_1 > p_2)/T \\ P_2 &= 1 - P_1, \end{aligned} \tag{6.9}$$

in two-group studies, and as

$$\begin{aligned} P_1 &= P\left[(p_1 > p_2) \cap (p_1 > p_3)|\theta_0, y, y^*, n, n^*\right] = \sum_{t=1}^{T} I\left[\bigcap_{i=2}^{3}(p_1 > p_i)\right]/T, \\ P_2 &= P\left[(p_2 > p_1) \cap (p_2 > p_3)|\theta_0, y, y^*, n, n^*\right] = \sum_{t=1}^{T} I\left[\bigcap_{i=1,\neq 2}^{3}(p_2 > p_i)\right]/T, \\ P_3 &= P\left[(p_3 > p_1) \cap (p_3 > p_2)|\theta_0, y, y^*, n, n^*\right] = \sum_{t=1}^{T} I\left[\bigcap_{i=1}^{2}(p_3 > p_i)\right]/T, \end{aligned} \tag{6.10}$$

in three-group studies. The predictive probabilities given in Eqs. 6.9 and 6.10 can then be incorporated in two- and three-group optimal designs in the same manner as the posterior efficacy comparisons.

6.4 Example 1: The Decreasingly Informative Prior Approach

This method was presented at BASS in 2013 and much of the following passages originally appeared in Sabo (2014). Lead-in and natural lead-in methods are designed to prohibit or constrain adaptation of allocation weights in early stages of a trial, when

estimates may be unreliable due to small sample sizes. Alternatively, one could use a posterior distribution to provide estimators that do not change much in early parts of a trial. Under the Bayesian framework, we could elicit decreasingly informative priors (DIP) which are mass or density functions that are functions of observed (n) and planned (N) sample sizes. These functions would also serve as skeptical priors in that they would be centered around some value θ_0 indicative of treatment equivalence when sample sizes are small. However, information is incrementally transferred to likelihood as n increases, making the prior decreasingly informative.

6.4.1 Decreasingly-Informative Prior Model

An alternative to the natural lead-in approach discussed in Thall and Wathen (2007) is the concept of a built-in lead-in component achieved by making the prior distributions functions of *non-accrued* patients. We first assume skeptical prior distributions for each treatment group by centering the efficacy rates around the same value p_0. To simultaneously keep the mode of the prior distribution at p_0 while also accounting for the accruing data, where $\pi()$ is the *common* distributional form of the priors for parameters p_j, $j = 1, \ldots, k$, we make these priors to be functions of the hypothesized value p_0 and the unobserved non-accrued subjects $N - n$ such that $\pi() = \pi(p_0, n, N)$, where N is the total planned sample size, and $n = \sum_{j=1}^{k} n_j$ is the total number of accrued patients.

Say we have binary outcomes in k groups and that we want to model those outcomes using the beta-binomial conjugate pair. Based on the general Bayesian set-up in Eq. 6.6, we could model outcomes in group j as $y_j \sim f(n_j, p_j) = \text{binomial}(n_j, p_j)$. The DIP for the group j success rate could be modeled as $p_j \sim \pi(p_0, n, N) = \text{beta}\,[1 + p_0(N - n), 1 + (1 - p_0)(N - n)]$, where the skeptical value p_θ is chosen as a single value or given its own hyperprior. This hyperprior could take any number of suitable forms, including $p_0 \sim \text{U}\,[\delta_1, \delta_2]$, where $0 \leq \delta_1 < \delta_2 \leq 1$ are suitably chosen upper and lower bounds for p_0, or even $p_0 \sim \text{beta}\,[1 + \delta_1, 1 + \delta_2]$ where δ_1 and δ_2 are chosen to elicit diffuse support for p_0. In either case, by parameterizing the priors with $a = 1 + p_0(N - n)$ and $b = 1 + (1 - p_0)(N - n)$, the desired mode is achieved

$$\text{mode} = \frac{a - 1}{a + b - 2} = \frac{p_0(N - n)}{p_0(N - n) + (1 - p_0)(N - n)} = p_0.$$

These prior distributions can be combined with likelihood functions for each treatment group to obtain posterior distributions for each parameter or a joint distribution of all parameters may be obtained. While using a hyperprior for p_0 may lead to a non-closed-form posterior, selecting a particular value for p_θ combined with *beta* priors and $binomial(n_i, p_i)$ likelihoods will lead to closed-form posterior distribution for the group j success rate $p_j \sim \text{beta}\,[1 + y_k + p_0(N - n), 1 + (n_j - y_j) + (1 - p_0)(N - n)]$. Regardless of the choices of prior and likelihood

and also between using posterior means, modes or efficacy comparisons, allocation weights are calculated using the optimal formulations found in Eqs. 6.1 and 6.2, not with Eqs. 6.3 and 6.4 since we are attempting to mimic the effect of a natural lead-in. At the beginning of a trial, the posterior estimates and probabilities depend only upon the skeptical prior information and are centered at the same value p_0, meaning that the allocation weights are equal. As more patients accrue into the trial, the prior information becomes increasingly less important relative to the accrued data. Thus, like the natural lead-in approach, the use of decreasingly-informative prior distributions forces the adaptation to move slowly during early parts of a trial and allows for more sensitive adaptation during latter parts of a trial.

6.4.2 Simulation Study for DIP Model

We performed a simulation study to compare the relative performance of Thall and Wathen's natural lead-in (TW) method with that of the decreasingly-informative prior (DIP) method of adaptive allocation in both two- and three-group trials. For the two-group case we assume that the first treatment has some superior true level of efficacy to the second treatment (i.e. $p_1 > p_2$), while in the three-group case we assume that $p_1 > p_2 > p_3$. In both cases we expect the first group of simulated patients to outperform those from the other groups, and thus expect both procedures to randomize more patients into the first treatment. For each new patient we simulate a random number $u \sim U[0, 1]$ to allocate between groups using Eqs. 6.1 (DIP) or 6.3 (TW) in the two-group case or Eqs. 6.2 (DIP) or 6.4 (TW) in the three-group case. The binary outcome for each patient is then probabilistically simulated based on a treatment-specific Bernouli distribution with success rate p_1, p_2 or p_3. The TW or DIP allocation ratios are then recalculated based on all currently available outcomes, and the process is repeated until the total number of patients is achieved, which is selected to attain at least 80% power in the balanced case.

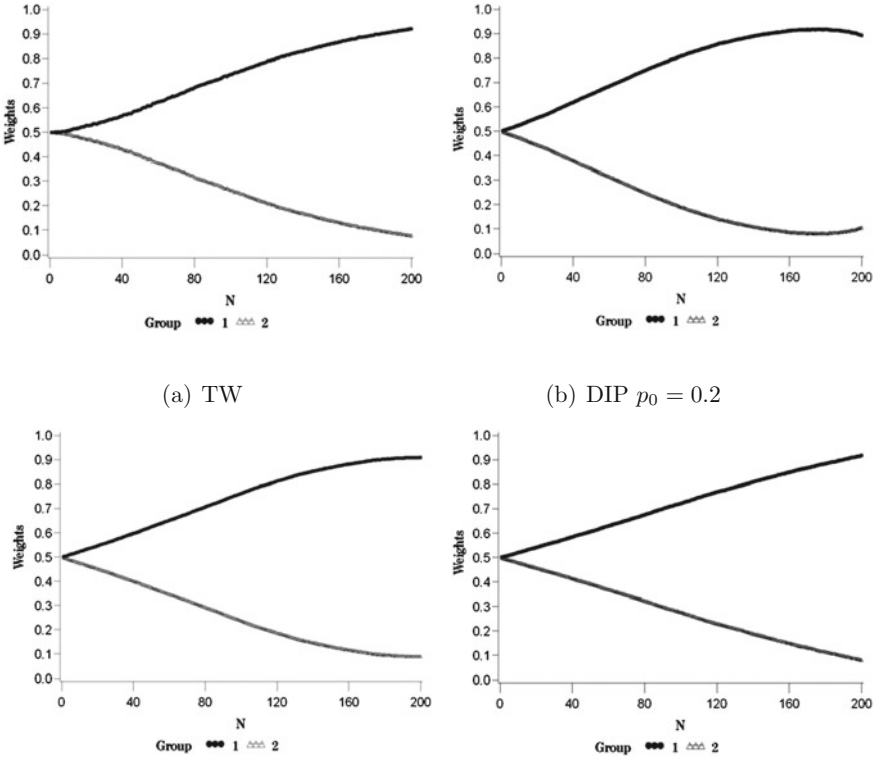
For the TW procedure we have assumed a non-informative $beta(1, 1)$ prior distribution on the efficacy proportion in each group. For the DIP procedure we examine situations where we select a particular prior value for p_0 and also where we select a non-informative hyperprior on that value. In the former case three values of p_0 are used to represent different realistic scenarios: one where we correctly guess the null hypothesized value, a second where we guess the null hypothesized value incorrectly by understating its value, and a third where we overstate its value. For the hyperprior case we select a diffuse and non-informative $U[0, 1]$ hyperprior in order to mimic the situation where we make no assumptions about the underlying efficacy about either group. We also investigate the use of either posterior means or posterior efficacy comparisons to calculate the allocation probabilities. Each trial was simulated 1000 times for each set of parameter values, from which we measure end-of-trial treatment-specific sample sizes (with standard deviations), empirical power, error rates, and allocation probabilities.

In Table 6.1 we see the results from two-group trials with a true effect-size of $\delta = 0.2$. In this case – which reflects overwhelming evidence of superiority for the first treatment – we see that the TW procedure maintains the highest power, though the DIP procedure is close when the pre-selected skeptical value p_0 is near the actual success rate in the second group. We also see that the methods provide similar allocation (in terms of final sample size), though the DIP method often does so with less variability than the natural lead-in approach. Figure 6.1 shows the average allocation probabilities for both groups throughout the trial. Here we see that the adaptation gradually increases with sample size, which is similar though not identical between the different approaches.

In Tables 6.2 and 6.3 we see comparisons in the two group case with a smaller effect size ($\delta = 0.15$) and where we now formulate the DIP procedure with a diffuse hyperprior. For both the TW and DIP methods we present the use of posterior means to calculate allocation weights in Table 6.2, and the use of efficacy comparisons in

Table 6.1 Simulation summaries for two group case (DIP with point mass). $*$ indicates correct choice of prior

True efficacy		$p_1 = 0.5$	$p_2 = 0.3$	$N = 200$
		DIP		
	TW	$p_0 = 0.2$	$p_0 = 0.3^*$	$p_0 = 0.4$
$\%(n_1 > n_2)$	99.7%	98.4%	99.4%	99.6%
Power	80.9%	71.8%	78.6%	79.7%
$\hat{n}_1$	144.7	153.8	148.3	143.0
$\hat{n}_2$	55.3	46.2	51.7	57.0
(SD)	(16.37)	(20.07)	(16.60)	(14.16)
True efficacy		$p_1 = 0.7$	$p_2 = 0.5$	$N = 200$
		DIP		
	TW	$p_0 = 0.4$	$p_0 = 0.5^*$	$p_0 = 0.6$
$\%(n_1 > n_2)$	98.9%	98.2%	98.5%	98.6%
Power	79.1%	75.8%	79.8%	77.5%
$\hat{n}_1$	143.6	150.0	146.8	142.9
$\hat{n}_2$	56.4	50.0	53.2	57.1
(SD)	(17.80)	(19.08)	(17.14)	(15.32)
True efficacy		$p_1 = 0.9$	$p_2 = 0.7$	$N = 200$
		DIP		
	TW	$p_0 = 0.6$	$p_0 = 0.7^*$	$p_0 = 0.8$
$\%(n_1 > n_2)$	99.9%	99.3%	99.8%	99.9%
Power	95.3%	92.2%	94.7%	93.8%
$\hat{n}_1$	153.0	154.0	153.0	151.5
$\hat{n}_2$	47.0	46.0	47.0	48.5
(SD)	(15.76)	(15.70)	(14.09)	(13.34)



(a) TW (b) DIP $p_0 = 0.2$

(c) DIP $p_0 = 0.3^*$ (d) DIP $p_0 = 0.4$

Fig. 6.1 Allocation probabilities for two group case (DIP with point mass). $*$ indicates correct choice of prior

Table 6.3. In the posterior mean case (Table 6.2) we see that though both methods provide some adaptation, neither meaningfully increases the expected number of successes from that achieved using balanced allocation. However, when posterior efficacy comparisons are used (Table 6.3), we see that in addition to providing more adaptation, both methods increase the expected number of treatment successes relative that achieved using balanced allocation.

Tables 6.4 and 6.5 present results from three-group trials using either posterior means and efficacy. In this case both posterior formulations provide increased treatment successes relative to balanced allocation. While the natural lead-in approach provides greater adaptation and more treatment successes, the DIP procedure has less variability in these measures.

Table 6.2 Simulation summaries for two group case (DIP with hyperprior; posterior mean)

True efficacy	$p_1 = 0.25$	$p_2 = 0.1$	$N = 200$
	Bal.	TW	DIP
Exp. Succ.	35.1 (3.85)	36.5 (4.28)	36.1 (4.00)
$\hat{n}_1$	100.2 (7.13)	110.6 (7.81)	105.3 (6.89)
$\hat{n}_2$	99.8 (7.13)	89.4 (7.81)	94.7 (6.89)
Power	80.0%	81.2%	80.3%
Error	0.0%	0.0%	0.0%
R_{50}	–	1.24 (0.15)	1.06 (0.04)
R_{75}	–	1.40 (0.22)	1.14 (0.07)
R_{100}	–	1.58 (0.30)	1.53 (0.25)
True efficacy	$p_1 = 0.55$	$p_2 = 0.4$	$N = 352$
	Bal.	TW	DIP
Exp. Succ.	167.2 (7.86)	167.9 (8.92)	167.8 (8.57)
$\hat{n}_1$	175.7 (9.47)	183.1 (10.23)	181.4 (9.65)
$\hat{n}_2$	176.3 (9.47)	168.9 (10.23)	170.6 (9.65)
Power	80.0%	81.3%	80.1%
Error	0.0%	0.0%	0.0%
R_{50}	–	1.08 (0.04)	1.05 (0.03)
R_{75}	–	1.13 (0.06)	1.10 (0.04)
R_{100}	–	1.17 (0.07)	1.17 (0.07)

6.5 Example 2: Accounting for Multiple Outcomes

There may be occasions when both the efficacy and toxicity of a novel treatment are under investigation, or where there are two important measures of efficacy. In such situations the meaning of a successful treatment could be defined as being one that is effective while not inducing toxicity, or is effective in more than one way. Investigators of such treatments may then want to utilize both outcomes in an outcome-adpative allocation process. One such method was presented at BASS in 2012, and much of the following passages appeared in Sabo et al. (2013).

6.5.1 Models for Dual Outcomes

We assume that the dual primary outcomes in the trial are dichotomous in nature (e.g. success or failure). The outcomes are not required to be immediately observable (though that definitely helps), provided that such delays are not too great with respect to the pace of patient enrollment and the planned duration of the trial Zelen (1969). At best, such delays merely prolong the period during which the original allocation

Table 6.3 Simulation summaries for two group case (DIP with hyperprior; posterior efficacy)

True efficacy	$p_1 = 0.25$	$p_2 = 0.1$	$N = 200$
	Bal.	TW	DIP
Exp. Succ.	35.1 (3.85)	40.6 (5.12)	38.5 (4.56)
$\hat{n}_1$	100.2 (7.13)	138.6 (12.85)	122.7 (7.60)
$\hat{n}_2$	99.8 (7.13)	61.4 (12.85)	77.3 (7.60)
Power	80.0%	77.8%	83.1%
Error	0.0%	0.0%	0.0%
R_{50}	–	2.56 (1.11)	1.44 (0.29)
R_{75}	–	5.51 (3.12)	2.24 (0.73)
R_{100}	–	12.2 (8.22)	12.3 (7.78)
True efficacy	$p_1 = 0.55$	$p_2 = 0.4$	$N = 352$
	Bal.	TW	DIP
Exp. Succ.	167.2 (7.86)	178.0 (15.11)	175.1 (12.79)
$\hat{n}_1$	175.7 (9.47)	244.1 (22.60)	226.9 (17.08)
$\hat{n}_2$	176.3 (9.47)	107.9 (22.60)	125.1 (17.08)
Power	80.0%	76.5%	81.5%
Error	0.0%	0.0%	0.0%
R_{50}	–	2.63 (1.12)	1.69 (0.46)
R_{75}	–	5.55 (3.14)	3.24 (1.57)
R_{100}	–	12.3 (8.44)	12.4 (8.13)

Table 6.4 Simulation summaries for three group case (DIP with hyperprior, true efficacy: $p_1 = 0.25$, $p_2 = 0.15$, $p_3 = 0.1$, $N = 345$, and $B = 0.2$)

Posterior mean					
	Bal.	Natural lead-in		DIP	
E(S)	57.2 (4.2)	62.6 (6.1)		59.0 (4.6)	
Power	79.5%	81.1%		78.5%	
Error	0.0%	1.1%		1.3%	
R_{50}	–	2.93 (2.17)	2.36 (1.79)	1.39 (0.75)	1.29 (0.69)
R_{75}	–	2.30 (1.05)	1.96 (0.98)	1.53 (0.77)	1.32 (0.70)
R_{100}	–	2.18 (0.75)	1.87 (0.73)	2.18 (0.74)	1.83 (0.72)
Posterior efficacy					
	Bal.	Natural lead-in		DIP	
E(S)	57.2 (4.2)	67.2 (6.5)		62.5 (5.6)	
Power	79.5%	77.0%		76.9%	
Error	0.0%	0.7%		0.9%	
R_{50}	–	3.68 (2.46)	3.32 (1.44)	1.88 (0.89)	1.81 (0.79)
R_{75}	–	3.07 (0.82)	2.99 (0.61)	2.43 (0.79)	2.37 (0.73)
R_{100}	–	2.93 (0.40)	2.94 (0.32)	2.91 (0.44)	2.93 (0.35)

Table 6.5 Simulation summaries for three group case (DIP with hyperprior, true efficacy: $p_1 = 0.55$, $p_2 = 0.45$, $p_3 = 0.4$, $N = 618$, and $B = 0.2$)

Posterior mean					
	Bal.	BS		DIP	
E(S)	288.4 (8.9)	294.4 (14.9)		290.2 (11.6)	
Power	78.8%	81.3%		80.1%	
Error	0.0%	0.7%		1.3%	
R_{50}	–	2.34 (1.50)	2.00 (1.58)	1.44 (0.70)	1.31 (0.67)
R_{75}	–	2.00 (0.84)	1.61 (0.87)	1.66 (0.67)	1.36 (0.65)
R_{100}	–	1.89 (0.61)	1.50 (0.63)	1.90 (0.58)	1.46 (0.61)
Posterior efficacy					
	Bal.	BS		DIP	
E(S)	288.4 (8.9)	305.6 (21.0)		302.4 (19.3)	
Power	78.8%	80.2%		80.3%	
Error	0.0%	0.5%		0.8%	
R_{50}	–	3.54 (2.26)	3.20 (1.39)	2.57 (0.76)	2.52 (0.72)
R_{75}	–	3.03 (0.68)	2.99 (0.51)	2.85 (0.52)	2.84 (0.47)
R_{100}	–	2.93 (0.38)	2.95 (0.29)	2.93 (0.40)	2.94 (0.32)

proportions are held constant, and at worst prohibit adaptation until latter stages of the trial, possibly even excluding changes all together (Berry and Eick 1995). The two outcomes do not need to be observed simultaneously in each patient; however, it must be noted that the algorithm would be biased in favor of the observed outcome in such cases. Further, we assume that the total sample size is fixed at some n, and that patients are randomized into one of k treatment groups or arms. This data will then be used to estimate θ_j and λ_j, $j = 1, \ldots, k$, where these parameters represent the mean of the first and second outcomes in each of the k treatments, respectively. Since we are assuming that our observations are dichotomous, these parameters would most likely represent proportions, but could be arranged to represent odds ratios or relative risks.

Bayesian methods can be used to turn the observed data and any beliefs concerning the two outcomes for each treatment into posterior probabilities on the k pairs of parameters in which we are interested. Regardless of how we calculate the posterior probabilities, or of what combinations we use for the two outcomes, we want the allocation weight for treatment j to be proportional to posterior probabilities of "positive" outcomes (e.g. efficacy), and proportional to the complements of "negative" outcomes (e.g. toxicity, futility). In the following subsections, we illustrate three different approaches for estimating allocation proportions. These approaches differ in how the posterior probabilities are calculated, based on whether we compare the outcome parameters directly between treatments or to hypothesized values.

6.5.1.1 Comparisons Between Treatment Arms

We first outline the case where we compare the "success" rates for both the first and second outcomes (θ_j and λ_j, respectively) for treatment j to the corresponding rates in all other treatments. The result of these comparisons are the posterior probabilities $P_{j\ell}^{\theta} = P(\theta_j > \theta_\ell)$ for the first outcome and $P_{j\ell}^{\lambda} = P(\lambda_j > \lambda_\ell)$ for the second outcome, where these comparisons are made for $\ell = 1, \ldots, k$, where $P_{jj}^{\theta} = P_{jj}^{\lambda} = 1$. If the θ_j and λ_j represent"positive" events (implying that larger values of $P_{j\ell}^{\theta}$ and $P_{j\ell}^{\lambda}$ indicate greater likelihoods of positive responses), then the allocation weight for the jth of k treatment arms is defined as

$$w_j = \frac{\left(\prod_{\ell=1}^{k} P_{j\ell}^{\theta} P_{j\ell}^{\lambda}\right)^{c(n)}}{\sum_{i=1}^{k}\left(\prod_{\ell=1}^{k} P_{i\ell}^{\theta} P_{i\ell}^{\lambda}\right)^{c(n)}},$$

where $c(n)$ is a suitably chosen tuning parameter that can adjust the pace of adaptation (Thall and Wathen 2007; Bello and Sabo 2016). Note that the allocation weight w_j for treatment j is proportional to the product of the posterior probabilities that the success rates for outcomes θ and λ in treatment j are greater than the success rates in every other treatment. Thus, the weight w_j can increase (or decrease) in a number of ways. For example, the allocation weight can increase if the success rate for just one of the outcomes in treatment j is larger than the corresponding rate in just one other treatment (assuming the probabilities for all other comparisons stay constant), or it could increase if treatment j has a higher outcome-one (or outcome-two) success rate than all other treatments; in this latter case the weight may increase more than in the former case. Conversely, w_j can decrease if treatment j is outperformed by another or several other treatments, with respect to outcome one, outcome two, or both.

Note that if one of the outcomes (say the second) were to represent a "negative" outcome (implying that higher rates for the λ_j represented undesirable outcomes, and that larger values of $P_{j\ell}^{\lambda}$ indicate a greater likelihood of that undesirable outcome happening), then we could simply focus on the"positive" complement $1 - P_{j\ell}^{\lambda}$ for each outcome in the allocation weight for the jth of k treatment arms.

6.5.1.2 Comparisons to Hypothesized Values

As mentioned in Huang et al. (2007), we could compare the "success" rates for each outcome in each treatment to hypothesized values (say p_o^{θ} and p_o^{λ}), should such values exist. For instance, we could compare the efficacy rates for a set of new treatments to a rate of 30% established by a "gold-standard" treatment, or physicians may wish to keep the toxicity rates below a 10% threshold. If such values are available, then the posterior probabilities $P_j^{\theta} = P(\theta_j > p_o^{\theta})$ and $P_j^{\lambda} = P(\lambda_j > p_o^{\lambda})$ can be calculated from the posterior distributions for each outcome in each treatment group. If we

assume that the two outcomes are "positively" valued, then the allocation weight for the jth of k treatment arms is defined as

$$w_j = \frac{\left(P_j^{\theta} P_j^{\lambda}\right)^{c(n)}}{\sum_{i=1}^{k} \left(P_i^{\theta} P_i^{\lambda}\right)^{c(n)}} \tag{6.11}$$

The weights described in Eq. 6.11 are proportional to the likelihood of positive outcomes *in single treatments*. While the treatments in this case are not directly compared with one another, the two outcomes in each group are compared to the same values. Treatments are thus indirectly compared, and superiority of one treatment over the hypothesized value will lead to an increased allocation weight for that treatment when either: such superiority is not as strong or lacking for other treatments, or those treatments are showing inferiority to the hypothesized values. The behavior of allocation weights for ambiguous scenarios would by their nature be difficult to predict.

6.5.1.3 Hybrid Approach

A likely scenario is the case where we want to compare one outcome between treatments and the other outcome within each treatment to a hypothetical standard. This could be the case if we wanted to determine the treatment with the greatest efficacy, provided that it kept toxicity below an allowable threshold. We assume that the first outcome is compared between treatments and the second is compared to a hypothesized value, so for each treatment j we will have $k-1$ posterior probabilities $P_{j\ell}^{\theta} = P(\theta_j > \theta_\ell)$, $\ell = 1, \ldots, k$ for the first outcome (recall $P_{jj}^{\theta} = 1$), and one posterior probability $P_j^{\lambda} = P(\lambda_j > p_o^{\lambda})$ for the second outcome. If we assume that both outcomes represent "positive" outcomes, then the allocation weight for the jth of k treatment arms is defined as

$$w_j = \frac{\left(P_j^{\lambda} \Pi_{\ell=1}^{k} P_{j\ell}^{\theta}\right)^{c(n)}}{\sum_{i=1}^{k} \left[P_i^{\lambda} \left(\Pi_{\ell=1}^{k} P_{i\ell}^{\theta}\right)\right]^{c(n)}}. \tag{6.12}$$

6.5.2 Simulation Study for Dual Objective Model

We calculate weights w_j for the $j = 1, \ldots, k$ treatment arms based assuming that posterior probabilities are raised to the power $(n/2N)$ as described in Eq. 6.3. By simulating $u \sim U[0, 1]$, we allocate the simulated patient to the j^{th} treatment arm if $\sum_{i=0}^{j-1} w_i < u < \sum_{i=1}^{j} w_i$, where $w_0 = 0$. At this point we simulate the efficacy and toxicity outcome for the new patient by generating a random outcome from a Bernoulli trial with efficacy probability $(p_e + \delta_j)$, where δ_j is the amount by which

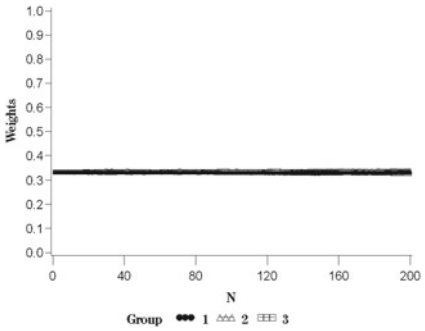
the probability of a successful outcome in the j^{th} treatment arm differs from p_e, and by also generating a random outcome from a second Bernoulli trial with toxicity probability $(p_t + \tau_j)$, where τ_j is the amount by which the probability of a toxic outcome in the j^{th} treatment arm differs from p_t. These new values are combined with the existing data to calculate posterior probabilities of both efficacious outcomes and toxic outcomes, which are in turn used to update the allocation weights, the method of which depends upon whether the performance of the treatment arms are being compared to hypothesized values, each other or both. One simulated clinical trial ends when the maximum sample size of $n = 200$ patients has been fully allocated. This process is repeated $m = 1000$ times for each set of assumed efficacy and toxicity rates.

Here we focus solely upon three-arm studies where efficacy is compared between arms and toxicity is compared to a hypothesized value. We assume informative and skeptical beta prior distributions for the P_j^e and P_j^t ($beta(1.3, 1.7)$ and $beta(1.1, 1.9)$, respectively). While the probability that a given treatment is less toxic than a hypothesized value ($P_t = 0.1$) can again be calculated using the posterior distribution of P_j^t, we use direct sampling to calculate $P_{jk}^e = P(P_j^e > P_k^e)$. Assuming treatment groups are independent, we simulate $m = 1000$ values each from the posterior distributions of the P_j^e, $j = 1, \ldots, k$, to obtain $(P_{1,j}^e, \ldots, P_{1000,j}^e)$ and estimate the posterior probability that treatment arm j is more successful than treatment arm k as

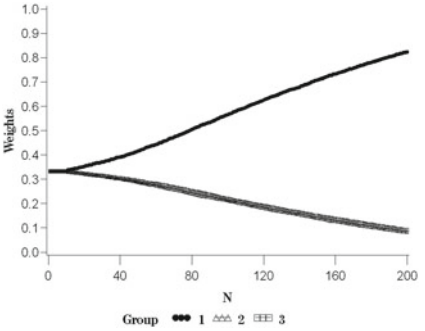
$$P_{jk} = P(P_j^e > P_k^e) = \frac{\sum_{i=1}^{m} I(P_{i,j}^e > P_{i,k}^e)}{1000},$$

where $I()$ is an indicator function.

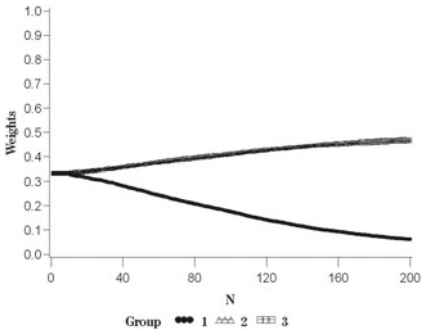
The average behaviors of allocation weights under various scenarios are found in Fig. 6.2. The first three panels show relatively straightforward scenarios, where (i) there is no efficacy or toxicity differences between the three treatments, (ii) the first treatment is more efficacious than the other two treatments, and (iii) the first treatment is more toxic than the other two treatments. The allocation weights do not change in the first case, skew in favor of the first treatment in the second case, and skew away from the first treatment in the third case. The average sample sizes presented in Table 6.6 for these three cases corroborate the visual results. In the ambiguous case where the first treatment is simultaneously more efficacious ($p_1^e = 0.5$) and toxic ($p_1^t = 0.2$) than the second and third treatments, Fig. 6.2 shows that the allocation weights change little during the trial, and the average numbers of subjects (Table 6.2) allocated between the three treatments (78.1, 60.1 and 61.7, respectively) are not as different as in first three cases. Here efficacy is slightly more meaningful than toxicity because in each treatment there are two inter-arm efficacy comparisons for every toxicity comparison. In the case where the first treatment is more efficacious ($p_1^e = 0.4$) than the second treatment ($p_2^e = 0.3$), which in turn is more efficacious than the third treatment ($p_3^e = 0.2$), more patients (101.0) are allocated to the first treatment than to the second (61.0) and third (38.0), with heavy favoring of treatment one resulting predominantly from the large efficacy difference between the first and



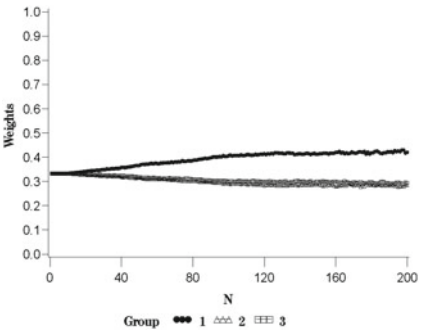
(a) $p_1^e = p_2^e = p_3^e = 0.3, p_1^t = p_2^t = p_3^t = 0.1$



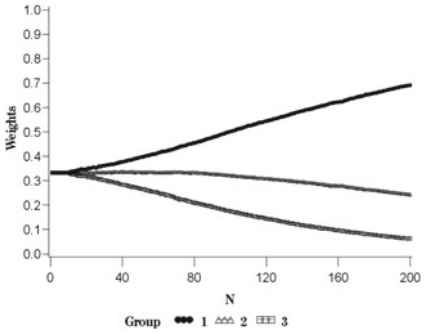
(b) $p_1^e = 0.5, p_2^e = p_3^e = 0.3; p_1^t = p_2^t = p_3^t = 0.1$



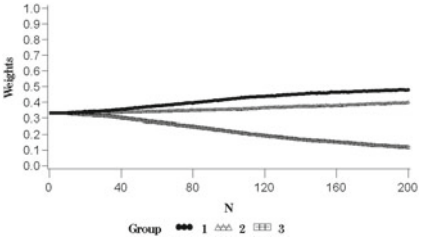
(c) $p_1^e = p_2^e = p_3^e = 0.3; p_1^t = 0.25, p_2^t = p_3^t = 0.1$



(d) $p_1^e = 0.5, p_2^e = p_3^e = 0.3; p_1^t = 0.2, p_2^t = p_3^t = 0.1$



(e) $p_1^e = 0.4, p_2^e = 0.3, p_3^e = 0.2; p_1^t = p_2^t = p_3^t = 0.1$



(f) $p_1^e = 0.4, p_2^e = 0.3, p_3^e = 0.2; p_1^t = 0.15, p_2^t = 0.1, p_3^t = 0.05$

Fig. 6.2 Average allocation weights based on number of accrued patients in 3 treatment arms for given efficacy and toxicity probabilities

Table 6.6 Average sample size (with standard deviation) for 3–arm trials: results from simulation study with $m = 1000$ repetitions with treatment comparisons made between treatments for efficacy and to hypothesized values for toxicity ($p_o^t = 0.1$)

Parameters	Sample size	Standard deviation	Parameters	Sample size	Standard deviation
$p_1^e = 0.3$			$p_1^e = 0.5$		
$p_2^e = 0.3$	$\hat{n}_1 = 66.0$	$SD_1 = 23.1$	$p_2^e = 0.3$	$\hat{n}_1 = 78.1$	$SD_1 = 28.4$
$p_3^e = 0.3$	$\hat{n}_2 = 66.2$	$SD_2 = 23.0$	$p_3^e = 0.3$	$\hat{n}_2 = 60.1$	$SD_2 = 24.4$
$p_1^t = 0.1$	$\hat{n}_3 = 67.8$	$SD_3 = 22.5$	$p_1^t = 0.2$	$\hat{n}_3 = 61.7$	$SD_3 = 25.5$
$p_2^t = 0.1$			$p_2^t = 0.1$		
$p_3^t = 0.1$			$p_3^t = 0.1$		
$p_1^e = 0.5$			$p_1^e = 0.4$		
$p_2^e = 0.3$	$\hat{n}_1 = 113.0$	$SD_1 = 21.9$	$p_2^e = 0.3$	$\hat{n}_1 = 101.0$	$SD_1 = 23.4$
$p_3^e = 0.3$	$\hat{n}_2 = 42.9$	$SD_2 = 17.1$	$p_3^e = 0.2$	$\hat{n}_2 = 61.0$	$SD_2 = 21.8$
$p_1^t = 0.1$	$\hat{n}_3 = 44.1$	$SD_3 = 17.3$	$p_1^t = 0.1$	$\hat{n}_3 = 38.0$	$SD_3 = 15.0$
$p_2^t = 0.1$			$p_2^t = 0.1$		
$p_3^t = 0.1$			$p_3^t = 0.1$		
$p_1^e = 0.3$			$p_1^e = 0.4$		
$p_2^e = 0.3$	$\hat{n}_1 = 37.2$	$SD_1 = 14.7$	$p_2^e = 0.3$	$\hat{n}_1 = 82.7$	$SD_1 = 26.0$
$p_3^e = 0.3$	$\hat{n}_2 = 81.2$	$SD_2 = 25.7$	$p_3^e = 0.2$	$\hat{n}_2 = 72.3$	$SD_2 = 24.0$
$p_1^t = 0.25$	$\hat{n}_3 = 81.6$	$SD_3 = 26.4$	$p_1^t = 0.15$	$\hat{n}_3 = 45.1$	$SD_3 = 18.7$
$p_2^t = 0.1$			$p_2^t = 0.1$		
$p_3^t = 0.1$			$p_3^t = 0.05$		

third treatments. For the other ambiguous case, where treatments one and two are sequentially more efficacious and toxic than treatment three, Fig. 6.2 shows that the weights turn against the third treatment in favor of the first and second (even though it is less toxic, it is also less efficacious than the other two). The weights for the first treatment are slightly higher than those for the second, and both are larger than the weights for the third treatment. The average number of total patients allocated to the first and second treatments (82.7 and 72.3, respectively) are also higher than the average number allocated to the third treatment (45.1). This is again due to the fact that while treatment two is less toxic than treatment one, treatment one is much more efficacious than treatment three. This might be a scenario where we consider different radical exponents for the two outcomes.

Using the same simulations from which the previous results were obtained, we have also calculated the percentage of simulations for which each of the three treatment arms had the highest number of allocated patients. These results are found in Table 6.7 and show that the most efficacious and least toxic treatments routinely receive the most patients. Also reported is the proportion of simulated trials (for both the adaptive and balanced allocation procedures) for which the various efficacy and toxicity rates were deemed significantly different between the three possible

Table 6.7 Percentage of larger samples and decisions in favor in 3–arm trials: results from simulation study with $m = 1000$ repetitions with treatment comparisons made between treatments for efficacy and to hypothesized values for toxicity ($p_o^t = 0.1$). Case 1: $p_1^e = p_2^e = p_3^e = 0.3$, $p_1^t = p_2^t = p_3^t = 0.1$. Case 2: $p_1^e = 0.5$, $p_2^e = p_3^e = 0.3$, $p_1^t = p_2^t = p_3^t = 0.1$. Case 3: $p_1^e = p_2^e = p_3^e = 0.3$, $p_1^t = 0.25$, $p_2^t = p_3^t = 0.1$. Case 4: $p_1^e = 0.5$, $p_2^e = p_3^e = 0.3$, $p_1^t = 0.2$, $p_2^t = p_3^t = 0.1$. Case 5: $p_1^e = 0.4$, $p_2^e = 0.3$, $p_3^e = 0.2$, $p_1^t = p_2^t = p_3^t = 0.1$. Case 6: $p_1^e = 0.4$, $p_2^e = 0.3$, $p_3^e = 0.2$, $p_1^t = 0.15$, $p_2^t = 0.1$, $p_3^t = 0.05$

Case comparison	Reject in Favor of		% of	Reject in Favor of		% of
	(Adapt) (%)	(Equal) (%)	Samples	(Adapt) (%)	(Equal) (%)	Samples
	Case 1			Case4		
Eff: 1v2	6.8	4.7	$n_1 > n_2, n_3$	77.2	76.6	$n_1 > n_2, n_3$
Eff: 1v3	6.1	5.1	31.2%	78.9	79.9	49.2%
Eff: 2v3	6.6	5.4	$n_2 > n_1, n_3$	8.0	4.7	$n_2 > n_1, n_3$
Tox: 1v2	7.1	4.4	33.9%	50.8	54.2	22.7%
Tox: 1v3	7.2	5.4	$n_3 > n_1, n_2$	52.7	49.9	$n_3 > n_1, n_2$
Tox: 2v3	7.3	6.4	33.8%	9.2	5.3	27.1%
	Case 2			Case5		
Eff 1v2	77.3	77.8	$n_1 > n_2, n_3$	32.7	32.4	$n_1 > n_2, n_3$
Eff 1v3	75.4	78.2	93.2%	74.7	82.0	79.7%
Eff 2v3	11.6	6.3	$n_2 > n_1, n_3$	36.7	40.2	$n_2 > n_1, n_3$
Tox: 1v2	4.3	5.8	3.2%	6.0	6.2	17.7%
Tox: 1v3	4.4	6.2	$n_3 > n_1, n_2$	7.5	6.7	$n_3 > n_1, n_2$
Tox: 2v3	11.6	5.6	3.3%	12.0	5.1	1.9%
	Case 3			Case6		
Eff 1v2	8.9	5.1	$n_1 > n_2, n_3$	35.3	34.7	$n_1 > n_2, n_3$
Eff 1v3	9.3	4.5	1.5%	81.8	82.9	55.0%
Eff 2v3	5.3	6.3	$n_2 > n_1, n_3$	38.2	43.0	$n_2 > n_1, n_3$
Tox: 1v2	75.3	78.0	48.7%	23.3	20.8	37.3%
Tox: 1v3	75.2	74.3	$n_3 > n_1, n_2$	58.6	61.8	$n_3 > n_1, n_2$
Tox: 2v3	4.9	4.8	48.8%	29.4	31.0	6.9%

treatment pairings (1 vs. 2, 1 vs. 3, and 2 vs. 3) using chi-square tests. The estimated proportions for the adaptive and fixed allocation methods are similar for Cases 1, 2, 3, 4 and 6, and the adaptive allocation method features a slight loss of power compared to the fixed allocation method in Case 5. These Cases show that the benefit of allocating subjects away from less efficacious or more toxic treatments may come at the cost of slightly lower power as compared to the fixed allocation method.

6.6 Discussion

Presented here are examples of adaptive allocation algorithms conducted under the Bayesian analytic framework. These methods – an adaptive allocation algorithm for dual outcomes, and the decreasingly informative prior approach – were originally presented at the BASS conference in 2012 and 2013, respectively. While these are emblematic of Bayesian techniques, they are by no means the only examples in the adaptive allocation literature. One particularly active research area is in covariate-adjusted response-adaptive allocation designs (Bandyopadhyay et al. 2007; Thall and Wathen 2007), where allocation algorithms can be balanced for patient characteristics, or where particular sub-groups can be given separate allocation weights. Another example is adaptive allocation designs for clinical trials with continuous Biswas and Bhattacharya (2016) our survival outcomes Zhang and Rosenberger (2007), which in general require entirely different algorithms and concepts of what constitutes "optimal" treatment outcomes.

References

Bandyopadhyay, U., Biswas, A., & Bhattacharya, R. (2007). A covariate adjusted two-stage allocation design for binary responses in randomized clinical trials. *Statists in Medicine*, *26*, 4386–4399.

Bello, G., & Sabo, R. T. (2016). Outcome-adaptive allocation with natural lead-in for three-group trials with binary outcomes. *Journal of Statistical Computation and Simulation*, *86*, 2441–2449.

Berry, D. A., & Eick, S. G. (1995). Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine*, *14*, 231–246.

Biswas, A., & Bhattacharya, R. (2016). Response-adaptive designs for continuous treatment responses in phase iii clinical trials: A review. *Statistical Methods in Medical Research*, *25*, 81–100.

Eisele, J. R. (1994). The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference*, *38*, 249–261.

Hu, F., & Zhang, L. X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *The Annals of Statistics*, *32*, 268–301.

Huang, X., Biswas, S., Oki, Y., Issa, J. P., & Berry, D. A. (2007). A parallel phase i/ii clinical trial design for combination therapies. *Biometrics*, *63*, 429–436.

Jeon, Y., & Hu, F. (2010). Optimal adaptive designs for binary response trials with three treatments. *Statistics in Biopharmaceutical Research*, *2*, 310–318.

Korn, E. L., & Freidlin, B. (2011). Outcome-adaptive randomization: Is it useful? *Journal of Clinical Oncology*, *29*, 771–776.

Rosenberger, W. F., Stallard, N., Ivanova, A., Haper, C. N., & Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics*, *57*, 909–913.

Sabo, R. T., co-authors. (2013). An outcome-adaptive allocation method for clinical trials with dual objectives. *Statistics in Biopharmaceutical Research*, *5*, 67–78.

Sabo, R. T. (2014). Adaptive allocation for binary outcomes using decreasingly informative priors. *Journal of Biopharmaceutical Statistics*, *24*, 569–578.

Sabo, R. T., & Bello, G. (2017). Optimal and lead-in adaptive allocation for binary outcomes: A comparison of methodologies. *Communications in Statistics: Theory and Methods*, *46*, 2823–2836.

Thall, P. F., & Wathen, J. K. (2007). Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine*, *24*, 1947–1964.

Thall, P. F., & Wathen, J. K. (2007). Practical bayesian adaptive randomization in clinical trials. *European Journal of Cancer*, *43*, 859–866.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*, 285–294.

Tymofyeyev, Y., Rosenberger, W. F., & Hu, F. (2007). Implementing optimal allocation in sequential binary response experiments. *Journal of the American Statistical Association*, *102*, 224–234.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistician*, *64*, 131–146.

Zhang, L., & Rosenberger, W. F. (2007). Response-adaptive randomization for survival trials: The parametric approach. *Journal of the Royal Statistical Society, Series C*, *34*, 562–569.

Chapter 7
Addressing High Placebo Response in Neuroscience Clinical Trials

Gheorghe Doros, Pilar Lim and Yuyin Liu

7.1 Background

7.1.1 Placebo Response in Major Depressive Disorder Trials

7.1.1.1 Factors Contributing to Placebo Response

In pharmacological research of Major Depressive Disorder (MDD), 38–50% of all short-term, acute, Phase 3 clinical studies failed to distinguish active drug from placebo, even for approved antidepressants (Gispen-de Wied et al. 2012; Khin et al. 2011), while 15% of the studies were considered negative (Chen et al. 2014). Although not all of these failed trials represent false negatives, meta-analyses of antidepressant studies suggest that high variability in the placebo response, independent of drug response, explains the majority of the variability in trial outcome (Alkermes announces advances 2014; Gispen-de Wied et al. 2012; Khin et al. 2011). The net result of this is an increasing number of failed clinical trials in depression, resulting in medications with potential clinical value not reaching patients in need. As summarized in a recent review by Rutherford et al. (2014), "High placebo response rates hamper efforts to detect signals of efficacy for new antidepressant medications, contributing to trial failures and delaying the delivery of new treatments to market",

G. Doros (✉) · Y. Liu
Department of Biostatistics, Boston University, 801 Massachusetts Avenue, Boston, MA 02118, USA
e-mail: doros@bu.edu

Y. Liu
e-mail: yuyin@bu.edu

P. Lim
Department of Quantitative Sciences, Janssen Research & Development, LLC, 1125 Trenton-Harbourton Road, Titusville, NJ 08560, USA
e-mail: plim@its.jnj.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_7

and by Gispen-de Wied et al. (2012), "All efforts should be made to optimize the clinical development of drugs in the psychiatric domain, in order to improve the intrinsic quality of studies and reduce the burden to both pharmaceutical companies and society of too many trials needed to complete the dossier."

There are numerous factors contributing to the problem of failed clinical trials in depression, including patient factors, investigator/site factors, and research design factors, discussed below.

Patient Factors Patient factors include expectancy of improvement (Meyer et al. 2002; Papakostas and Fava 2009; Rutherford and Roose 2013; Sotsky et al. 1991), past experiences with antidepressant medications, the patient's perception of the informed consent process, and the patient's relationship with the investigator. Patients' expectancy of improvement in clinical trials of antidepressants is thought to be a primary mechanism of placebo responding (Rutherford and Roose 2013). There are several factors which contribute to expectancy bias: (1) the informed consent procedure, during which patients are informed about the study design, past effectiveness of the drug under study, side effects associated with the study drug, and the investigator's opinion/biases regarding potential effectiveness of the study drug, (2) past experience with antidepressant treatment, and (3) the probability of receiving active drug versus placebo. Meta-analyses suggest that the design of a clinical trial, especially the probability of receiving active drug versus placebo, influences patients' expectancy of clinical improvement regardless of treatment assignment, which has a negative impact on signal detection. Recent studies that have measured patient expectancy at baseline report that higher expectancy predicts greater symptom improvement over the course of treatment (Krell et al. 2004; Meyer et al. 2002; Rutherford et al. 2013).

Another patient factor that can contribute to placebo response is clinical presentation, which includes severity and duration of the depressive episode (Fournier et al. 2010; Kirsch et al. 2008; Stein et al. 2006), subtype of depressive disorder, and comorbidities (particularly anxiety disorders). In addition, the natural course of psychiatric illnesses, involving spontaneous improvement and worsening of patients' symptoms unrelated to study treatment, is an uncontrolled source of variability. Other patient factors influencing placebo response include psychiatric history, (including family history of psychiatric disorders), use of concomitant medications (Wernicke et al. 1997), positively perceived therapeutic effects of undergoing medical procedures, and the methods used for patient recruitment (source and types of patient recruited into trials).

Investigator/Site Factors Investigator/site factors include experience conducting clinical trials, rater bias, and the therapeutic setting at the site.

Rater bias, reflecting a conscious or subconscious determination to detect effects of the study medication, can influence ratings of symptom severity (Landin et al. 2000; Mundt et al. 2007; Rief et al. 2009). The rater's guess of treatment assignment to rate patients according to a preconceived, desired outcome in placebo-controlled depression trials has been shown to impact ratings, with a negative rating bias against demonstrating improvement when the rater believes that the subject is on placebo,

and a positive rating bias demonstrating improvement when the rater believes that the subject is on active drug treatment (Chen et al. 2015).

The effect of the therapeutic setting (supportive contact with investigative site staff) also impacts the placebo response. In a meta-analysis of 41 randomized, controlled antidepressant trials for major depression, the greater the number of study visits (and therefore, greater amount of potentially therapeutic contact with site staff), the higher the placebo response rate (Posternak and Zimmerman 2007). Although patients receiving active drug also demonstrated more improvement with an increasing number of study visits, the relative effect of two extra study visits was approximately 50% greater in the placebo group than in the active treatment group, suggesting that greater therapeutic contact has a differential effect on placebo response.

Research Design Factors Finally, research design factors play a significant role in placebo response, and, being under the control of the sponsor, are the most amenable to modification to reduce the placebo response. These design factors include validity and responsiveness of the primary outcome measure, duration of the trial, flexible versus fixed dose designs, number of treatment arms, and number of trial sites that can each impact the placebo response even after controlling for similar study population, inclusion criteria, study site/geographical region, test product, and similar study design.

Certain psychiatric rating scales have been shown to be more consistently responsive to change in pharmacotherapy trials, and therefore the choice of primary outcome measure is important to ensure adequate ability to detect an efficacy signal (Carmody et al. 2006; Khan et al. 2002, 2004). Measurement factors represent a source of bias and error, and are significant given the subjective scales employed in psychiatric trials. Regression to the mean, a statistical event that occurs when repeated measurements associated with random error are made on the same trial participant over time, is another source of apparent, but not true, symptom change in antidepressant trials (Rutherford and Roose 2013).

Shorter trial duration is associated with higher antidepressant trial success rates (6-week versus 8-week trials) (Khin et al. 2011) and, in general, fixed dose studies have a slightly higher success rate than flexible dose trials (Khan et al. 2003; Khin et al. 2011). Regarding number of treatment arms, mean placebo response rate has been shown to be lower in studies including 1 active treatment versus 2 or more active treatments (Khan et al. 2004). In fact, as the probability of receiving placebo decreases, the placebo response rates increase (Papakostas and Fava 2009). For each 10% decrease in the probability of receiving placebo, the probability of response to the active drug increased by 1.8%, while the probability of response to placebo increased by 2.6%. When comparing drug response between placebo-controlled trials and active comparator trials (2 or more drugs with no placebo group), patient expectancy has been shown to influence treatment response: mean drug response rates in active comparator trials were greater than in placebo-controlled trials, a finding that is consistent across studies of children, adolescents, adults, and older adults (Rutherford 2009; Sneed et al. 2008). In general, the smaller the number of trial sites, the less variability is introduced, and the higher the trial success rate (Bridge et al. 2009; Robinson and Rickels 2000).

In summary, the high rate of placebo response/variability is a major factor contributing to the failure of clinical trials in psychiatry, and particularly, clinical trials in MDD. The growing number of failed trials has made developing psychiatric medications increasingly more time consuming and expensive when compared with non-CNS indications (Cressey 2011; Nutt and Goodwin 2011).

7.1.1.2 Methods of Controlling for Placebo Response

Patient Expectancy Since patient expectancy of improvement is considered to be a prime driver of placebo response and decreased placebo-drug differences, researchers have implemented a number of strategies to reduce expectancy. One strategy is to implement a placebo lead-in period to identify and exclude subjects who respond rapidly to placebo treatment. This approach, in theory, addresses both patient expectancy bias and nonspecific therapeutic effects of the health care setting. However, analyses of studies with single-blind placebo lead-in periods have not found this design element to be beneficial in improving signal detection (Rutherford et al. 2011; Stein et al. 2006; Trivedi and Rush 1994). A single, published study incorporating a double-blind lead-in period (both study personnel and patients were blind to the duration of the placebo lead-in) showed some benefit in signal detection (Douglas et al. 2001).

Rater Bias/Measurement Error Several methods may be employed to reduce rater bias, measurement error, and regression to the mean. Dual assessments, centralized ratings, rater drift monitoring, etc. are techniques frequently being used in trials. Raters can be blinded as to the timing of the baseline assessment when the severity score required for randomization is assessed, and different rating scales can be used to determine subject eligibility and to serve as the primary outcome measure in the planned analyses.

Functional unblinding, occurring when a subject reports side effects caused by the study drug, may increase the subject's and investigator's expectancy of improvement, resulting in measurement bias. This can be mitigated through the use of an independent rater blinded to the treatment assignment who is unaware of any reported side effects.

7.1.1.3 Clinical Trial Design

As discussed above, several design features have been shown to contribute to reducing the risk of a failed clinical trial in MDD, including: a higher probability of receiving placebo; enrichment with higher baseline severity of illness; fewer study visits (reducing therapeutic contact); a single active treatment arm; and fewer study sites (reducing placebo variability).

In addition to these measures, the use of novel trial designs aimed at addressing expectancy bias and untangling the complex relationship between true medication effect and the contribution of nonspecific, "placebo" effects to overall medication

response will be important in addressing the issue of the growing placebo response rate in clinical trials of antidepressants. Trial designs implementing double-blind lead-in periods, in which the randomization timepoint is blinded to both subject and investigator, and trials implementing enrichment designs such as double randomizations, in which subjects not responding to placebo in the first period of the study are re-randomized to study drug or placebo, are increasingly being used to determine true drug response. Equal randomization to active compound or higher allocation to placebo may reduce subjects' expectancy of improvement.

As a result, various trial design features for reducing the risk of a failed study have been developed, including the implementation of placebo lead-in periods and reduction in the number and complexity of study visits. Despite the implementation of such design features, failed clinical trials remain a major issue hampering the development of urgently-needed new therapies for MDD.

7.1.2 Novel Study Designs to Address Placebo Response

The most commonly employed design used in antidepressant clinical trials today, the parallel-group design, has been used for the past 30+ years (Khan et al. 2003) with only minor modifications and has suffered high rates of failed trials with high placebo response felt to be a major contributing factor (Khin et al. 2011; Thase 1999). No single intervention or combination of interventions has been shown to be successful in eliminating placebo response. Given that the contributions from the multiple factors discussed above cannot be completely eliminated, recent novel trial designs such as the sequential parallel design (SPD) and the doubly-randomized delayed start (DRDS) design are being employed to increase the efficiency of placebo-controlled trials using the concepts of re-randomization and enrichment (Chen et al. 2011; Liu et al. 2012). This is the first significant advance in clinical trial methodology in the last quarter-century.

7.1.2.1 Sequential Parallel Design

The SPD is a two-period design that was proposed in 2003 by Dr. Maurizio Fava and colleagues of the Massachusetts General Hospital (Fava et al. 2003) to address the emerging issue of the placebo response and to reduce sample size in psychiatric clinical trials. The original SPD randomizes all subjects only once, at the beginning of Period 1, into 3 treatment sequences: treatment-placebo, placebo-treatment, and placebo-placebo; this trial design was used in the TRD-2 and ADAPT-A trials for MDD (Study of 6(S)-5-MTHF 2014; Fava et al. 2012; Papakostas et al. 2012).

Statistical Analysis Researchers from academic institutions, the United States Food and Drug Administration, and the pharmaceutical industry have published many papers on analytical methods for SPD data. Because the SPD involves two periods and assumes that subjects who have failed to respond to placebo in the first

period are even less likely to respond to placebo in the second period, Period 2 reveals a higher drug-placebo difference. Efficient statistical tests aggregate data from both periods while controlling the Type I error. Statistical tests for the SPD using binary measurements include the likelihood ratio test, Wald test, score test, and combination test. (Fava et al. 2003; Huang and Tamura 2010; Ivanova et al. 2011; Tamura et al. 2011) For continuous measurements, the combination test was studied but the design with no re-randomization at Period 2 poses technical analysis difficulties.

Current Applications So far, at least 21 trials have used or plan to use the SPD (Fava et al. 2003). Some trials are funded by the United States National Institute of Health while others are sponsored or collaborated by institutions (eg, Massachusetts General Hospital, University of Connecticut, Vanderbilt University, Yale University) and pharmaceutical companies (eg, Alkermes, Bristol-Myers Squibb, Pfizer). In the United States, the Food and Drug Administration has recently approved the use of SPD in three Phase 3 pivotal studies for compounds in development for the treatment of MDD (Alkermes announces advances 2014).

7.2 Binary Response

Fava et al. (2003) developed a test for used in the SPCD that is based on linear combination of treatment differences in the two stages. Fava et al. (2003) and Tamura and Huang (2007) concluded that the SPCD is more efficient than a two-arm placebo-controlled single-stage design under a broad range of assumptions. Ivanova et al. (2011) proposed a 1 and 2 degree of freedom (DOF) score tests for treatment effect in SPCD trials. In this section, we will discuss these analysis methods for binary responses in SPCD trials.

7.2.1 Original Method

Fava et al. (2003) proposed a test for binary responses when they brought up the concept of SPCD trials. They assume that patients will be randomized into three groups. The first two groups will initially receive placebo, those patients that do not respond to placebo will receive placebo (group 1) or drug (group 2). The third group will initially receive drug. The proportions randomized to the three groups will be a, a, and $(1 - 2a)$.

Let p_1, q_1 be the response rates to the first administration of drug and placebo respectively and let p_2, q_2 be the responses to the second treatment. To analyze these data, they use a statistic based on $\delta_\omega = \omega(p_1 - q_1) + (1 - \omega)(p_2 - q_2)$. The weight, ω and the randomization fraction, a are chosen to maximize the power of the test, based on the alternative hypothesis.

Table 7.1 Notations of the original method

Group	Response	Frequency	Probability
1 Placebo-placebo	No-Yes	n_{11}	$(1-q_1)q_2$
	No-No	n_{12}	$(1-q_1)(1-q_2)$
	Yes	n_{13}	q_1
2 Placebo-Drug	No-Yes	n_{21}	$(1-q_1)p_2$
	No-No	n_{22}	$(1-q_1)(1-p_2)$
	Yes	n_{23}	q_1
3 Drug	Yes	n_{31}	p_1
	No	n_{32}	$(1-p_1)$

The standard error for δ_ω requires a special formula because some of the same patients who are included in the estimation of p_2, and q_2 are included in the estimation of p_1, and q_1. The delta method was used to compute the standard error of δ_ω. The computation is facilitated by considering by considering the following table of outcomes, where in this case p_1, p_2, q_1, and q_2 are the theoretical probabilities rather than the observed relative frequencies (Table 7.1).

Then

$$\delta_\omega = \omega\left(\frac{n_{31}}{n(1-2a)} - \frac{(n_{13}+n_{23})}{2na}\right) + (1-\omega)\left(\frac{n_{21}}{(n_{21}+n_{22})} - \frac{n_{11}}{(n_{11}+n_{12})}\right)$$

where n is the total number of patients. Let Q be the column vector of derivatives of δ_ω with respect to $n_{31}, n_{13}, n_{23}, n_{21}, n_{22}, n_{11}, n_{12}$, from the multinomial distribution. Then the standard error of δ_ω is given by $sqrt(Q'VQ)$, which is computed with the observed values of p_1, p_2, q_1, and q_2. The formula below is a simplified calculation of the standard error of δ_ω:

$$D = -2(-1+2a)p_2(-1+\omega)^2 + 2(-1+2a){p_2}^2(-1+\omega)^2$$

$$-2(-1+2a)q_2(-1+\omega)^2 + 2(-1+2a){q_2}^2(-1+\omega)^2$$

$$+(-1+q_1)((-1+q_1)q_1 + 2a(-p_1 + {p_1}^2 + q_1 - {q_1}^2))\omega^2$$

$$N = 2a(-1+2a)(-1+q_1)$$

$$s = \sqrt{\frac{D}{N}}$$

To test the null hypothesis, they use $Z = \delta_\omega / s$. The values of a and ω were calculated by substituting the alternative hypothetical values of p_1, p_2, q_1, and q_2 and find the values of a and ω that maximize Z. The power of the test is then $\Phi(Z - 1.96)$, where Φ is the cumulative distribution of the normal distribution.

7.2.2 Score Tests

Ivanova et al. (2011) developed a 1 and 2 degrees of freedom (DOF) score tests for treatment effect in SPCD trials. The 1 DOF test uses a test parameter r equal to a known ratio between treatment differences in Stage II and Stage I. The 2 DOF test does not require assumptions about the relationship between the treatment differences in the two stages. It is likely that some subjects who participated in Stage I will not participate in Stage II; therefore, all formulae have been developed to accommodate the possibility of such dropouts.

Description of the Design Let the total sample size in the trial be n with n_1 subjects in the placebo-placebo group, n_2 subjects in the placebo-drug group, and n_3 subjects in the drug-drug group, $n_1 + n_2 + n_3 = n$, with $n_1 = n_2$ and $a = n_1/n$. Therefore, subjects are randomized to the three groups according to an $a : a : (1 - 2a)$ ratio. Because $a = n_1/n$ and $n_1 = n_2$, in theory, the range for a is $0 \leq a \leq 0.5$, with $a = 0.5$ corresponding to a two-stage design with all patients receiving placebo in Stage I, and $a = 0.25$ corresponding to a two-stage design with equal allocation to drug and placebo in the first stage. Denote p_1 = Pr(Drug Response in Stage I), q_1 = Pr(Placebo Response in Stage I), p_2 = Pr(Drug Response in Stage II|Placebo Non-responder in Stage I), and q_2 = Pr(Placebo Response in Stage II|Placebo Non-responder in Stage I). The SPCD is depicted in Table 7.2. In the placebo-placebo group, n_{11} is the observed number of non-responders in Stage I who respond in Stage II, n_{12} is the observed number of non-responders in both stages, n_{13} is the observed number of responders in Stage I, and n_{14} is the number of non-responders who dropped out after Stage I, $n_1 \geq n_{11} + n_{12} + n_{13}$. Similarly, in the placebo-drug group, n_{21} is the observed number of non-responders in Stage I who respond in Stage II, n_{22} is the observed number of non-responders in both stages, n_{23} is the observed number of responders in Stage I, and n_{24} is the number of non-responders who dropped out after Stage I, $n_2 \geq n_{21} + n_{22} + n_{23}$. In the drug-drug group, n_{31} and n_{32} are the numbers of non-responders and responders, respectively, in Stage I.

The Score Test with 1 DOF Let s be the probability that a first-stage placebo non-responder continues to the second stage. Because subjects are independent, it follows that

$$n_{11} + n_{12} \sim Bin(n_1 - n_{13}, s) n_{21} + n_{22} \sim Bin(n_2 - n_{23}, s)$$

The dropout process is assumed random and independent of future outcomes. The joint likelihood for (p_1, q_1, p_2, q_2, s) based on n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{31}, and n_{32} is then

$$L_0(p_1, q_1, p_2, q_2, s) \propto {p_1}^{n_{32}}(1 - p_1)^{n_{31}} q_1^{n_{23}+n_{13}}(1 - q_1)^{n_1+n_2-n_{23}-n_{13}} {p_2}^{n_{21}}(1 - p_2)^{n_{22}}$$

$$\times {q_2}^{n_{11}}(1 - q_2)^{n_{12}} s^{n_{11}+n_{12}+n_{21}+n_{22}}(1 - s)^{n_1+n_2-n_{13}-n_{23}-n_{11}-n_{12}-n_{21}-n_{22}}$$

Table 7.2 Notations of score test

Treatment		Response			
Period 1	Period 2	Period 1	Period 2	Count	Probability
Placebo	Placebo(n_1)	No	Yes	n_{11}	$s(1-q_1)q_2$
		No	No	n_{12}	$s(1-q_1)(1-q_2)$
		Yes	.	n_{13}	q_1
		No	Missing	n_{14}	$(1-s)(1-q_1)$
Placebo	Drug(n_2)	No	Yes	n_{21}	$s(1-q_1)p_2$
		No	No	n_{22}	$s(1-q_1)(1-p_2)$
		Yes	.	n_{23}	q_1
		No	Missing	n_{24}	$(1-s)(1-q_1)$
Drug	Drug(n_3)	No	.	n_{31}	$1-p_1$
		Yes	.	n_{32}	p_1

Define treatment effects $\Delta_1 = p_1 - q_1$, $\Delta_2 = p_2 - q_2$ and $\Delta_2 = \rho\Delta_1$, the ratio of treatment effects. The test is derived under the assumption that ρ is known. They refer to r, $r = \rho$, as a test parameter to distinguish it from ρ. They restrict r to be $0 \le r < +\infty$. The parameters (p_1, q_1, p_2, q_2, s) are transformed to (Δ_1, q_1, q_2, s). Then, $p_1 = \Delta_1 + q_1$, $p_2 = r\Delta_1 + q_2$, and the re-parameterized likelihood is

$$L_1(\Delta_1, q_1, q_2, s) \propto (\Delta_1 + q_1)^{n_{32}}(1 - q_1 - \Delta_1)^{n_{31}}$$

$$\times q_1^{n_{23}+n_{13}}(1-q_1)^{n_1+n_2-n_{23}-n_{13}}(r\Delta_1 + q_2)^{n_{21}}(1 - r\Delta_1 - q_2)^{n_{22}}$$

$$\times q_2^{n_{11}}(1-q_2)^{n_{12}}s^{n_{11}+n_{12}+n_{21}+n_{22}}(1-s)^{n_1+n_2-n_{13}-n_{23}-n_{11}-n_{12}-n_{21}-n_{22}}.$$

The null hypothesis of interest is testing $H_0 : \Delta_1 = 0$. Under H_0, $\Delta_1 = p_1 - q_1 = 0, (p_2 - q_2) = r(p_1 - q_1) = r\Delta_1 = 0$, and therefore H_0 implies that treatment effects are 0 in both stages.

Maximum likelihood estimates under H_0 obtained by setting $\Delta_1 = 0$ and solving the likelihood equations for q_1, q_2, and s are as follows:

$$\tilde{q_1} = \frac{n_{32} + n_{13} + n_{23}}{n},$$

$$\tilde{q_2} = \frac{n_{11} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}},$$

$$\tilde{s} = \frac{n_{11} + n_{12} + n_{21} + n_{22}}{n_1 + n_2 - n_{13} - n_{23}}.$$

The test statistic is

$$T_1 = \frac{(\frac{n_{32}}{\tilde{q_1}} - \frac{n_3 - n_{32}}{1-\tilde{q_1}} + \frac{rn_{21}}{\tilde{q_2}} - \frac{rn_{22}}{1-\tilde{q_2}})^2}{\frac{n_3(n_1+n_2)}{\tilde{q_1}(1-\tilde{q_1})(n_1+n_2+n_3)} - \frac{r^2(1-\tilde{q_1})n_1n_2\tilde{s}}{\tilde{q_2}(1-\tilde{q_2})(n_1+n_2)}} \tag{7.1}$$

The asymptotic distribution of T_1 under H_0 is Chi-squared with 1 DOF. The test is similar in concept to that in Fava et al. (2003), where a weight ω is chosen to combine data from Stage I and II. If $r = 0$, the test is equivalent to the score test that uses data from Stage I only. If r is large, T_1 is close to the score test statistic that uses data from Stage II only. In the theoretical extreme case, if $a = 0.5$, that is $n_3 = 0$,

$$T_1 = \frac{\tilde{q_2}(1-\tilde{q_2})(n_1+n_2)}{(1-\tilde{q_1})sn_1n_2}\left(\frac{n_{21}}{\tilde{q_2}} - \frac{n_{22}}{1-\tilde{q_2}}\right)^2 \tag{7.2}$$

This test does not depend on r and is a score test for Stage II data only. Generally, as a increases, the choice of r has less effect on the power of the test.

The Score Test with 2 DOF In the 2 DOF test, the constraint regarding relationship between treatment effects Δ_1 and Δ_2 no longer exists. Changing parameter from (p_1, q_1, p_2, q_2, s) to (Δ_1, Δ_2, q_1, q_2, s), the likelihood is

$$L_2(\Delta_1, \Delta_2, q_1, q_2, s) \propto (\Delta_1 + q_1)^{n_{32}}(1 - q_1 - \Delta_1)^{n_{31}}$$

$$\times q_1^{n_{23}+n_{13}}(1-q_1)^{n_1+n_2-n_{23}-n_{13}}(\Delta_2 + q_2)^{n_{21}}(1 - \Delta_2 - q_2)^{n_{22}}$$

$$\times q_2^{n_{11}}(1-q_2)^{n_{12}}s^{n_{11}+n_{12}+n_{21}+n_{22}}(1-s)^{n_1+n_2-n_{13}-n_{23}-n_{11}-n_{12}-n_{21}-n_{22}}.$$

The hypothesis testing: H_0: $\Delta_1 = \Delta_2 = 0$. The score test statistic is

$$\begin{aligned} T_2 = {} & \frac{\tilde{q_1}(1-\tilde{q_1})(n_1+n_2+n_3)}{(n_1+n_2)n_3}\left(\frac{n_{32}}{\tilde{q_1}} - \frac{n_3 - n_{32}}{1-\tilde{q_1}}\right)^2 \\ & + \frac{\tilde{q_2}(1-\tilde{q_2})(n_1+n_2)}{(1-\tilde{q_1})\tilde{s}n_1n_2}\left(\frac{n_{21}}{\tilde{q_2}} - \frac{n_{22}}{1-\tilde{q_2}}\right)^2 \end{aligned} \tag{7.3}$$

The distribution of T_2 under H_0 is Chi-squared with 2 DOF. When $a = 0.5$, $T_2 = T_1$ in Equation (7.2). Because T_1 is compared against 1 DOF Chi-square and T_2 is compared against 2 DOF Chi-square, the 2 DOF test would have lower power than the 1 DOF test when $a = 0.5$.

7.3 Continuous Response

In most depression trials, it is also of interest to analyze continuous efficacy data such as change from baseline to the end of study in a total score such as the Hamilton Depression Rating Scale (HDRS). The data from each phase of the study have its own unique set of dependent and independent variables. How would one analyze such data from a sequential parallel comparison design (SPCD)? Many statistical approaches have been brought out in the past 10 years. In this chapter, we will discuss the following four analysis methods: seemingly unrelated regression (SUR), ordinary least square (OLS) estimation, repeated measures model (RMM) and weighted repeated measures model (WRMM).

In this work, we consider the SPCD format with one active treatment and placebo in Phase I and re-randomization of the Phase I placebo non-responders to active treatment or placebo in Phase II. Without much loss of generality, we can also assume that, in Phase I, the subjects are randomized 2 : 1 to placebo versus active treatment, while in the Phase II, placebo non-responders are randomized 1 : 1 to placebo versus active treatment. Placebo responders continue on placebo in Phase II, whereas subjects randomized to active treatment in Phase I continue on active treatment in Phase II. The parametrization of the design is presented in the flowchart in Fig. 7.1.

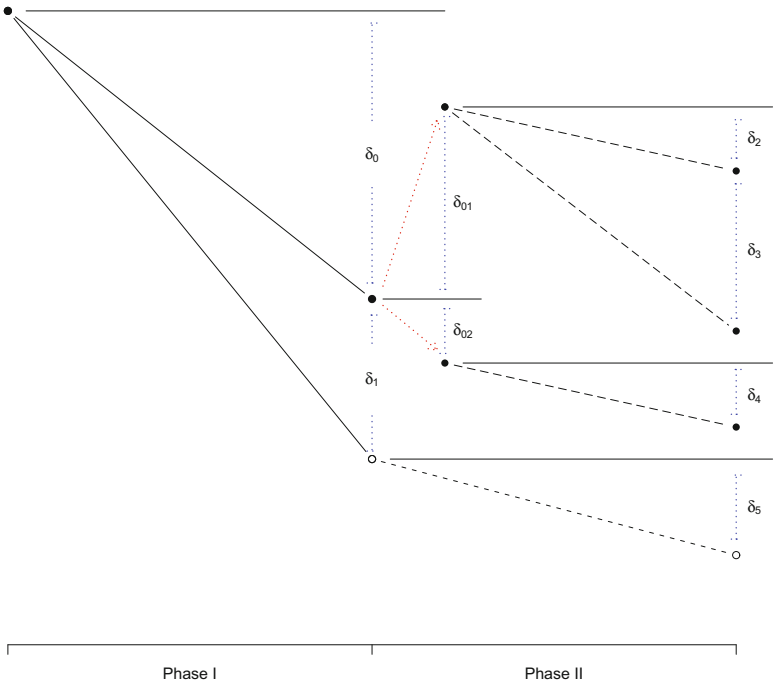


Fig. 7.1 Treatment effect diagram in sequential parallel comparison design trial

In this figure

- δ_0 denotes the mean change from baseline to the end of Phase I for subjects in the placebo group.
- δ_1 denotes the difference in mean change from baseline to the end of Phase I between subjects in the active treatment group and subjects in the placebo group (treatment effect in Phase I).
- δ_{01} denotes the difference between the mean outcome at the end of Phase I for all placebo subjects and mean outcome for placebo non-responders. It can be viewed as the amount of deviation from the change in outcome for all placebo subjects incurred by placebo non-responders.
- δ_{02} denotes the difference between the mean outcome at the end of Phase I for all placebo subjects and mean outcome for placebo responders. It can be viewed as the amount of deviation from the change in outcome for all placebo subjects incurred by placebo responders.
- δ_2 denotes the mean change from the baseline of Phase II to the end of Phase II for placebo non-responders who were randomized to placebo in Phase II.
- δ_3 denotes the difference in mean change from the baseline of Phase II to the end of Phase II between placebo non-responders subjects randomized to active treatment and placebo non-responders randomized to placebo in Phase II (treatment effect in Phase II).
- δ_4 denotes the mean change from the baseline of Phase II to the end of Phase II for placebo responders.
- δ_5 denotes the mean change from the baseline of Phase II to the end of Phase II for subjects in the active treatment group in Phase I.

Under the assumption of a normally distributed outcome, for a full specification of the design, each parameter would need to be specified along with the variance-covariance parameters. The parameters δ_0, δ_{01}, δ_{02}, δ_2 and δ_4 are the characteristics of subjects who receive only placebo, thus the information can be elicited from previous trials. The parameter δ_2 represents the placebo response among placebo non-responders, while δ_4 represents the placebo response among placebo responders. Both can be elicited from historical data.

The parameter δ_5 represents the Phase II treatment response for patients randomized to active treatment in Phase I. This parameter must be elicited. The parameters δ_1 and δ_3 are the treatment effect in Phase I and Phase II, respectively, and the overall treatment effect is defined as a weighted average of the two,

$$\delta_\omega = \omega\delta_1 + (1 - \omega)\delta_3$$

The magnitude of the treatment effect in Phase I, δ_1, treatment effect in Phase II, δ_3, and the weight w, are the parameters that mainly determine the size of the trial. In the following sections, we assume a 12-week trial, with Phase I baseline at Week 0, the end of Phase I/beginning of Phase II at Week 6, and the end of Phase II at Week 12. The four analysis approaches will be introduced with the consideration and determination of these parameters.

7.3.1 Seemingly Unrelated Regression

Tamura and Huang (2007) used the seemingly unrelated regression to analyze the SPCD data by taking into account the intra-patient correlation between efficacy data of placebo non-responders from both stage. They assume the data from the two phases of the study can be expressed via a linear model:

$$\Delta Y_{i1} = \alpha_{01} + \alpha_{11} Y_{i1,0} + \delta_1 G_{i1} + \epsilon_{i1}; \quad i = 1 : N$$

$$\Delta Y_{i2} = \alpha_{02} + \alpha_{12} Y_{i2,0} + \delta_3 G_{i2} + \epsilon_{i2}; \quad i = 1 : n_{NR}$$

where $\Delta Y_{i1} = Y_{i6} - Y_{i1,0}$ is the difference between the Week 6 and baseline outcome scores, $Y_{i1,0}$ is the Week 0 (Phase I baseline) outcome score and G_{i1} is the active treatment indicator for the data during Phase I; $\Delta Y_{i2} = Y_{i12} - Y_{i2,0}$ is the difference between the Week 12 and Week 6 outcome scores, $Y_{i2,0}$ is the Week 6 (Phase II baseline) outcome score and G_{i2} is the active treatment indicator for the data during Phase II for placebo non-responders.

It is assumed that the variances are constant from patient to patient within each phase; however, for a patient with data from both phases of the study the random errors from the two phases of the study may be correlated.

It is also assumed that patients are independent, $E(\boldsymbol{\epsilon_i}) = 0$, and that the within patient residual vector has a variance covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{21} \end{pmatrix}$$

If Σ were known, the generalized least squares solution (GLS) would be available; however in practice Σ is unknown and an estimator of Σ must be used. The analogue to the GLS using the estimator of Σ is sometimes called the estimated generalized least squares estimator (EGLS). As long as the estimator of Σ is consistent, the EGLS will have the sam asymptotic distribution as the GLS. Most software packages use the estimator of Σ based on the ordinary least squares residuals from the analysis of each phase separately.

In this application, SUR ignores the occurrence of missing Phase II data from placebo responders. SUR differs from a repeated measures analysis which treats the three groups as fixed covariates. In the repeated measures analyses, the model would be assumed to be true even for the population of patients with missing data.

Consider the situation where treatment is the only factor in both Phase I and Phase II models; δ_1 represents the treatment effect in Phase I, and δ_3 represents the treatment effect in placebo non-responders in Phase II, and the null hypothesis of no treatment effect is:

$$H_0 : \delta_1 = \delta_3 = 0.$$

It is assumed that treatment effect in each period would be in the same direction and consider a weighted average test statistic:

$$Z_{SUR} = \frac{\omega\hat{\delta_1} + (1-\omega)\hat{\delta_3}}{\sqrt{\omega^2 Var(\hat{\delta_1}) + 2\omega(1-\omega)Cov(\hat{\delta_1}, \hat{\delta_3}) + (1-\omega)^2 Var(\hat{\delta_3})}} \tag{7.4}$$

where ω is between 0 and 1. If $\sigma_{12} = 0$, then seemingly unrelated regression is equivalent to ordinary least squares on each phase of the study separately. In this case, the estimates of treatment effect in either phase are the simple treatment mean differences. Zellner (1962) showed that by considering the correlation and the information on explanatory variables in both phases simultaneously, the EGLS will be more efficient than the phase by phase ordinary least squares estimates.

7.3.2 Ordinary Least Square Estimation

Chen et al. (2011) examined the covariance between the ordinary least square test statistics for two stages of Fava's SPCD. For simplicity, they did not consider missing outcome data and the selected the two-sample t-test to estimate the treatment effect in each stage. The OLS estimates of δ_1 and δ_3 can be written as

$$\hat{\delta_1} = \frac{1}{N/3}\sum_{i=2N/3+1}^{N} Y_{i6} - \frac{1}{2N/3}\sum_{i=1}^{2N/3} Y_{i6}$$

and

$$\hat{\delta_3} = \frac{1}{n_{NR}/2}\sum_{j=n_{NR}/2+1}^{n_{NR}} Y_{j12} - \frac{1}{n_{NR}/2}\sum_{j=1}^{n_{NR}/2} Y_{j12}$$

respectively, where subjects are randomized 2:1 to placebo versus active treatment at Phase I, and are re-randomized 1:1 to placebo versus active treatment at Phase II for placebo non-responders. N is the total number of subjects enrolled in this trial, and n_{NR} is the number of placebo non-responders entering Phase II. Y_{i6} denotes the outcome measure for the i-th subject at Week 6 (the end of Phase I), and Y_{j12} denotes the outcome measure for the j-th subject at Week 12 (the end of Phase II). Assuming a constant correlation between the continuous endpoint measures in two phases for placebo non-responders entering Phase II and a constant variance of the continuous endpoint measures for all subjects in each phase, it is analytically proved that the covariance between $\hat{\delta_1}$ and $\hat{\delta_3}$ equals 0, and the weighted OLS statistic

$$Z_{OLS} = \frac{\omega\hat{\delta_1} + (1-\omega)\hat{\delta_3}}{\sqrt{\omega^2 Var(\hat{\delta_1}) + (1-\omega)^2 Var(\hat{\delta_3})}} \tag{7.5}$$

can be used as alternative to Z_{SUR} to test the null hypothesis $H_0 : \delta_1 = \delta_3 = 0$. Based on this finding, the asymptotic power of the weighted OLS statistic is equal to

$$\Phi\left(-1.96 - \frac{\omega\delta_1 + (1-\omega)\delta_3}{\sqrt{\frac{9\omega^2{\sigma_1}^2}{2N} + \frac{4(1-\omega)^2{\sigma_2}^2}{n_{NR}}}} \right)$$

It is also analytically proved that when an analysis of covariance (ANCOVA) is applied to each stage of the design, the unconditional covariance between the two adjusted treatment effects is also zero asymptotically. Therefore, the aforementioned weighted OLS statistic can also be applied to the SPCD when there is a need to incorporate a covariate into analysis of variance.

7.3.3 Repeated Measures Model

Both SUR and OLS methods ignore some of the data collected during the trial, e.g. the Phase II data on those on active treatment or who responded to placebo in Phase I are simply ignored in the OLS setting and treated as missing in the SUR setting. Doros et al. (2013) proposed a model that incorporates all the trial data and, using simulations, they demonstrate that, under a wide range of scenarios, this methodology preserves the type I error without compromising power.

At the core of this method is a repeated measure statistical model that uses all available data. This model estimates the treatment effect by using the data collected at baseline, end of Phase I and end of Phase II. Still, assume a 12-week trial, with Phase I baseline at Week 0, the end of Phase I/beginning of Phase II at Week 6 and the end of Phase II at Week 12.

The Equations Under The Model: This model consists of four equations.

1. An equation relating the outcome at the end of Phase I [defined as change in score from baseline (Week 0) to the end of Phase I (Week 6)] to the outcome at baseline and treatment allocation during the first 6 weeks for the subjects on placebo and on active treatment in the study

$$\Delta Y_{i1} = \alpha_{01} + \alpha_{11} Y_{i1,0} + \delta_1 G_{i1} + \epsilon_{i1}; \quad i = 1 : N,$$

where $\Delta Y_{i1} = Y_{i6} - Y_{i1,0}$ is the difference between the end of Phase I (Week 6) and baseline (Week 0) outcome scores, $Y_{i1,0}$ is the baseline outcome score and G_{i1} is the active treatment indicator for the data during Phase I.

2. An equation relating the outcome at the end of Phase II [defined as the change in the outcome from Week 6 to Week 12 (end of Phase II)] to Phase II baseline and treatment allocation for placebo non-responders

$$\Delta Y_{i2} = \alpha_{02} + \alpha_{12} Y_{i2,0} + \delta_3 G_{i2} + \epsilon_{i2}; \quad i = 1 : n_{NR},$$

where $\Delta Y_{i2} = Y_{i12} - Y_{i2,0}$ is the difference between the end of Phase II (Week 12) and the baseline of Phase II (Week 6) outcome score, $Y_{i2,0}$ is the Phase II baseline outcome score and G_{i2} is the treatment indicator for active treatment for the data during Phase II for placebo non-responders.

3. An equation relating the outcome at the end of Phase II to Phase II baseline for placebo responders

$$\Delta Y_{i2} = \alpha_{03} + \alpha_{13} Y_{i2,0} + \epsilon_{i3}; \quad i = (n_{NR} + 1) : (n_{NR} + n_R),$$

where ΔY_{i2} is the difference between the end of Phase II (Week 12) and the baseline of Phase II (Week 6) outcome score, $Y_{i2,0}$ is the Phase II baseline outcome score for placebo responders.

4. An equation relating the outcome at the end of Phase II and Phase II baseline for subjects randomized to active treatment in Phase I

$$\Delta Y_{i2} = \alpha_{04} + \alpha_{14} Y_{i2,0} + \epsilon_{i4}; \quad i = (n_{NR} + n_R + 1) : N.$$

Above N is the total number of subjects enrolled in Phase I, n_{NR}, n_R, and $n_T = N - n_{NR} - n_R$ are the respective numbers of placebo non-responders, placebo responders, and subjects on active treatment, who continue to Phase II.

The Covariance Under The Model: The error terms $\{\epsilon_{i1}\}_i$, $\{\epsilon_{i2}\}_i$, $\{\epsilon_{i3}\}_i$, and $\{\epsilon_{i4}\}_i$ are assumed to be independent and identically distributed across individuals. Since the subjects contributing data to the estimation carried out in Phase I also contribute data to the estimation in Phase II, the errors corresponding to data from the same subjects should be correlated. To achieve this, the following assumptions are made:

$$(\epsilon_{i1}, \epsilon_{i2}) \sim N\left[(0,0), \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right]; i = 1 : n_{NR},$$

$$(\epsilon_{i1}, \epsilon_{i3}) \sim N\left[(0,0), \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right]; i = (n_{NR} + 1) : (n_{NR} + n_R),$$

$$(\epsilon_{i1}, \epsilon_{i4}) \sim N\left[(0,0), \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right]; i = (n_{NR} + n_R + 1) : N.$$

Thus, the same correlation matrices are assumed for subjects who contribute data to the efficacy evaluation in Phase II and for subjects who do not.

Of note, there are three properties of this model that set it aside from the previous approaches:

1. All the data is used in estimating the parameters of the above model.
2. Different mean structure for Phase II outcome is assumed for the model in placebo responders and subjects on active treatment in Phase I.
3. Mean and variance parameters of the model are estimated then using the restricted maximum likelihood, thus, resulting in asymptotically efficient estimates.

The contrast of interest

$$\delta_\omega = \omega\delta_1 + (1-\omega)\delta_3$$

which represents the weighted average treatment effect in all subjects in Phase I and in placebo non-responders in Phase II, is estimated by $\hat{\delta}_\omega = \omega\hat{\delta}_1 + (1-\omega)\hat{\delta}_3$, with $\hat{\delta}_1$ and $\hat{\delta}_3$ are the model based estimates of δ_1 and δ_3, respectively. A test for $H_0: \ \omega\delta_1 + (1-\omega)\delta_3 = 0$ is based on the test statistic

$$Z_{RMM} = \frac{\omega\hat{\delta}_1 + (1-\omega)\hat{\delta}_3}{\sqrt{\omega^2 Var(\hat{\delta}_1) + 2\omega(1-\omega)Cov(\hat{\delta}_1, \hat{\delta}_3) + (1-\omega)^2 Var(\hat{\delta}_3)}} \tag{7.6}$$

where the variances and covariances are estimated based on the model above. This test statistic is of the same form as (7.4), the difference being the model used in estimating the parameters and the variance-covariance estimates of these estimators. Under the null hypothesis H_0, the above test statistic is assumed to follow a standard normal distribution, thus, allowing us to carry out the test and obtain p-values.

7.3.4 Weighted Repeated Measures Model

Rybin et al. (2015) proposed a change in the method of analysis of SPCD trial data that includes more subjects into the estimation of the Phase II effect by using all re-randomized Phase I placebo subjects and the weighted estimation for the Phase II effect. This would hopefully increase precision of the estimate and increase the test power. Figure 7.2a shows the graphical representation on the effects estimated with the repeated measure model. The blue arrows represent outcome progression in the placebo group and the red arrows - in the active treatment group. Still, the parameter of interest

$$\delta_\omega = \omega\delta_1 + (1-\omega)\delta_3$$

is the weighted average treatment effect in all subjects in Phase I and in placebo non-responders in Phase II.

The placebo response can be defined explicitly in terms of the change of the outcome measure from baseline in subjects treated with placebo. However, simple

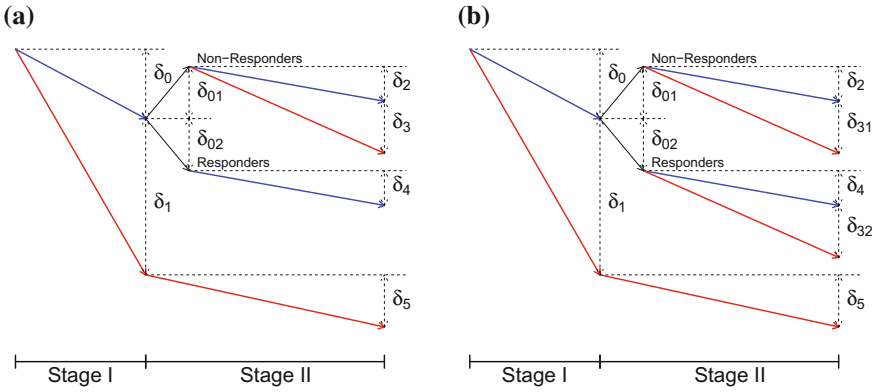


Fig. 7.2 Parametrization

classification of subjects as placebo responders and placebo non-responders based on the change is subject to misclassification. In general, this method tries to define placebo response as a characteristic that is, to some extent, present in each subject of the trial.

If that characteristic is known or measurable, it can be scaled to range from 0 to 1 so that scores close to 0 correspond to high placebo response, and scores close to 1 correspond to low placebo response. These can be seen as subject contributions (weights) in the placebo response corrected treatment effect in Phase II of the SPCD in Phase I placebo subjects. In previous approaches proposed by Tamura and Huang, and Chen et al. placebo non-responders are used to estimate treatment effect in Phase II, while placebo responders are excluded from the estimation, which is tantamount to assigning placebo non-responders a weight of 1 and placebo responders a weight of 0 for the analysis of Phase II data. Thus, the 0/1 classification can be seen as a particular case of this more general approach. It is assumed that the non-response characteristic is known.

The parameterization of this approach is presented on Fig. 7.2b. Both placebo responders and placebo non-responders are re-randomized after Phase I in a 1:1 ratio to placebo and active treatment in Phase II. Therefore, the treatment effect in Phase II can be computed in all Phase I placebo subjects. Figure 7.2b notations are the same as Fig. 7.2a notations with one exception: the Phase II treatment effect can no be defined in both placebo non-responders (indicated by δ_{31}) and placebo responders (indicated by δ_{32}).

The Equations Under The Model: This model consists of three equations.

1. An equation relating the outcome at the end of Phase I to the outcome at baseline and treatment allocation during the first 6 weeks for the subjects on placebo and on active treatment in the study

$$\Delta Y_{i1} = \alpha_{01} + \alpha_{11} Y_{i1,0} + \delta_1 G_{i1} + \epsilon_{i1}; \quad i = 1 : N,$$

where $\Delta Y_{i1} = Y_{i6} - Y_{i1,0}$ is the difference between the end of Phase I (Week 6) and baseline (Week 0) outcome scores, $Y_{i1,0}$ is the baseline outcome score and G_{i1} is the active treatment indicator for the data during Phase I.

2. An equation relating the outcome at the end of Phase II to Phase II baseline (Week 6 outcome) and the new treatment assignment for re-randomized Phase I placebo group.

$$\Delta Y_{i2} = \alpha_{02} + \alpha_{12} Y_{i2,0} + \delta_3 G_{i2} + \epsilon_{i2}; \quad i = 1 : n_P,$$

where $\Delta Y_{i2} = Y_{i12} - Y_{i2,0}$ is the difference between the end of Phase II (Week 12) and the baseline of Phase II (Week 6) outcome scores, $Y_{i2,0}$ is the Phase II baseline outcome score and G_{i2} is the treatment indicator for active treatment for the data during Phase II for Phase I placebo group.

3. An equation relating the outcome at the end of Phase II and Phase II baseline for subjects randomized to drug in Phase I

$$\Delta Y_{i2} = \alpha_{03} + \alpha_{13} Y_{i2,0} + \epsilon_{i3}; \quad i = n_P + 1 : N.$$

Above N is the total number of subjects enrolled in Phase I, n_P, and n_T are the respective numbers of Phase I placebo and Stage I treatment subjects.

The Covariance Under The Model: The errors ϵ under the model are distributed normally with mean $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 \Sigma$, where σ^2 is unknown, and Σ is defined as follows, reflecting correlation between Phase I and Phase II (ρ_{12}):

$$\Sigma_i = \mathbf{w_i}^{-1/2} \begin{bmatrix} 1 \rho_{12} \\ \rho_{12} 1 \end{bmatrix} \mathbf{w_i}^{-1/2}; \quad i = 1 : n_P$$

$$\Sigma_i = \begin{bmatrix} 1 \rho_{12} \\ \rho_{12} 1 \end{bmatrix}; \quad i = n_P + 1 : N$$

The weights are set to 1 for all subjects in Phase I. In Phase II, all Phase I treated subjects are assigned weight 1 but Phase I placebo subjects are assigned weights based on their non-response to placebo. Therefore, $\mathbf{w_i}$ takes the following form:

$$\mathbf{w_i} = \begin{bmatrix} 1 & 0 \\ 0 & w_i \end{bmatrix}$$

In matrix form the model can be written as $\Delta Y_i = X_i \beta + \epsilon_I$, where δY_i is a vector of outcome measures and X_i is the covariate matrix for individual i. The generalized least squares estimate for coefficients is

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} (X_i' \Sigma_i^{-1} X_i) \right\}^{-1} \sum_{i=1}^{N} X_i' \Sigma_i^{-1} Y_i$$

With Σ_i known, the variance of the estimate is

$$Var(\hat{\beta}) = \sigma^2 \left\{ \sum_{i=1}^{N} (X_i' \Sigma_i^{-1} X_i) \right\}^{-1}$$

The estimate is unbiased, and for a given $\mathbf{w_i}$, both σ^2 and Σ_i are estimated from the data using restricted maximum likelihood. To account for possible model misspecification of the variance, the 'Sandwich Estimator' of the variance, a robust estimate of the variance, can be used:

$$\tilde{Var}(\hat{\beta}) = \left\{ \sum_{i=1}^{N} (X_i' \Sigma_i^{-1} X_i) \right\}^{-1} \sum_{i=1}^{N} (X_i' \Sigma_i^{-1} V_i \Sigma_i^{-1} X_i) \left\{ \sum_{i=1}^{N} (X_i' \Sigma_i^{-1} X_i) \right\}^{-1}$$

where $V_i = (\Delta Y_i - X_i \hat{\beta})(\Delta Y_i - X_i \hat{\beta})'$. Therefore, the treatment effect in Phase I $\hat{\delta_1}$ and the treatment effect in Phase II can be estimated with corresponding variances and covariance $Var(\hat{\delta_1})$, $Var(\hat{\delta_3})$, and $Cov(\hat{\delta_1}, \hat{\delta_3})$.

The Treatment Effect Estimate: The treatment effect is defined as a linear combination of the treatment effect in Phase I δ_1 and the treatment effect in Phase II δ_3 estimated in all Phase I placebo subjects. The weight for SPCD treatment effect ω is specified. A test for $H_0: \ \delta_\omega = \omega\delta_1 + (1-\omega)\delta_3 = 0$ is based on the test statistic

$$Z_{WRMM} = \frac{\omega\hat{\delta_1} + (1-\omega)\hat{\delta_3}}{\sqrt{\omega^2 Var(\hat{\delta_1}) + 2\omega(1-\omega)Cov(\hat{\delta_1}, \hat{\delta_3}) + (1-\omega)^2 Var(\hat{\delta_3})}}, \quad (7.7)$$

where the effects, variances and covariances are estimated with the model above. It is assumed that Z_{WRMM} to follow approximately standard normal distribution under the null hypothesis.

In reality, the placebo non-response characteristic is an unknown quantity. Two ways are proposed for defining and estimating this characteristic.

Prediction of placebo non-response prediction: Placebo response can be affected by a number of factors such as age, gender, and severity of disease. Therefore, it is appropriate to predict the probability of responding to placebo based on a subject's characteristics.

They propose to generate a subject's propensity score of placebo response (or inversely placebo non-response) based on the subject's characteristics. A simple prediction can be based solely on baseline outcome measure (and hence on the baseline disease severity), as presented below.

$$w(y_{i0}) = pr(R_i = 0 | Y_{i,0} = y_{i,0}); \quad i = 1 : n_P$$

Above $Y_{i,0}$ is the outcome measure at baseline for subject i, R_i is a response indicator taking value 0 or 1 (for non-response or response to placebo, respectively) and n_P is the number of Phase I placebo subjects. In practice $w(y_{i0})$ can be easily estimated with logistic regression applied to Phase I placebo subjects data.

Placebo non-response as a characteristic based on trial data: Alternatively, one can measure subject's non-response to placebo based on the data observed in Phase I. In general, each subject can be placed in the R^n space of the n measured characteristics. Because the characteristics are selected to express placebo non-response (actual change from baseline, percent change from baseline, disease severity etc.), the subjects with similar degree of non-response will naturally be closer to each other. And hence the subject's coordinates can be used as a measure of non-response. They propose to use K-means clustering to determine subjects relative positions in the R^n space. K is set to 2 in order to group 'placebo responders' and 'placebo non-responders'.

In this example, two variables are considered for clustering: percent change from baseline in Phase I and the baseline value in Phase II. It is expected that non-responders would have low percent change from baseline in Phase I and a high baseline value in Phase II. The former quality is simply by definition of non-response, and the latter quality is due to the smaller change in Phase I and the inverse relationship of the disease severity and placebo response. These two variables are likely to be correlated. Therefore, to preserve relative distances and to base the analysis on the Euclidean distance they propose to compute two principal components for the two measures previously mentioned. Then, following procedure can be used to determine individual placebo response measure.

1. Perform K-means clustering (with $K = 2$) on the two principal components. The centers of clusters, the variability within clusters and the total variability are retrieved from the analysis.
2. The center-point coordinates (adjusted for within cluster variability) are computed as follows:

$$c_1 = \frac{m_{11}s_{21} + m_{21}s_{11}}{s_{11} + s_{21}}$$

$$c_2 = \frac{m_{12}s_{22} + m_{22}s_{12}}{s_{12} + s_{22}},$$

   where c_1 and c_2 are X and Y center coordinates, m_{11}, m_{21} are mean X coordinates of cluster 1 and cluster 2, m_{12} and m_{22} are the mean Y coordinates of the clusters, s_{11} and s_{21} are the standard deviations of the X coordinates, and s_{12} and s_{22} are the standard deviations of Y coordinates.
3. The distance d_i to the center-point for subject i is computed as follows:

$$d_i = (-1)^{C_i}\sqrt{(p_{1i} - c_1)^2 + (p_{2i} - c_2)^2}; \quad i = 1 : n_P,$$

where $C_i \subset \{1, 2\}$ is the cluster (in this example, cluster 1 subjects have greater change in Phase I and decreased baseline value in Phase II (responders), and cluster 2 subjects have decreased change in Phase I and greater baseline value in Phase II (non-responders)), and p_{1i} and p_{2i} are the two principal components for ith subject.

4. The subject specific scores are then computed as follows:

$$w_{i,k} = \Phi_k(d_i)$$

where Φ_k is the CDF function of normal distribution with mean 0 and standard deviation $k \times TSD$ (TSD is total standard deviation determined in K-means clustering step). The parameter k regulates values close to the tails of the distribution. Lower values of k produce the w function that is close in shape to a step-function, and higher values of k produce function that is close to linear.

This approach can be easily generalized for larger number of measured parameters. The number of principal components may change accordingly.

Both methods provide us with subject-specific measures of placebo non-response, ranging from 0 to 1. The scores close to 0 correspond to high placebo response and the scores close to 1 correspond to low placebo response. These can be seen as subject contributions (weights) in the placebo response corrected analysis.

7.4 ADAPT-A Trial Example

The multi-center, double-blind placebo-controlled study of the efficacy of low-dose aripiprazole (2 mg/day) adjunctive to antidepressant therapy (ADT) in the treatment of major depressive disorder patients with a history of inadequate response to prior ADT (ADAPT-A) was conducted using SPCD (Fava et al. 2009, 2012). After screening, patients were randomized to either aripiprazole 2 mg/day ($n = 54$), placebo-placebo ($n = 83$) or placebo-aripiprazole ($n = 84$) with a 2:3:3 ratio. The patients were followed for 60 days (30 days Phase I and 30 days Phase II). The key secondary endpoint was the difference in absolute change from baseline in the Montgomery-Asberg Depression Rating Scale (MADRS) score between aripiprazole 2 mg and placebo. The non-response was defined at the end of Phase I as less than a 50% decrease in MADRS total score from baseline and a MADRS score greater than 16. The summaries of the outcome are presented in Table 7.3.

Prior antidepressant therapy data showed that Phase I outcome changes were either negatively or not significantly correlated with Phase II changes. Pearson correlation was $\rho = -0.32 (p = 0.027)$ for aripiprazole-aripiprazole group, $\rho = -0.08 (p = 0.503)$ for placebo-placebo group, and $\rho = -0.18 (p = 0.119)$ for placebo-aripiprazole group.

Table 7.3 Baseline MADRS score and change for phase I and phase II (ADAPT-A Trial)

Time	Measure	Aripiprazole	Placebo
Baseline	N	54	167
	Mean ± SD	30.69 ± 4.02	31.20 ± 4.75
Stage I	N	52	162
	Mean ± SD	−8.46 ± 7.18	−8.26 ± 8.15
Stage II	N	58	61
	Mean ± SD	−5.84 ± 6.98	−3.30 ± 6.00

7.4.1 Binary Response

The Original Method: The sample sizes in the three groups were $n_1 = 83$, $n_2 = 84$, $n_3 = 54$, and the observed counts in the trial were $n_{11} = 5$, $n_{12} = 58$, $n_{13} = 15$, $n_{21} = 11$, $n_{22} = 50$, $n_{23} = 14$, $n_{31} = 44$, $n_{32} = 10$. There were 138 non-responders to placebo in Stage I, 124 of them participated in Stage II, yielding the estimated retention rate $\tilde{s} = 124/138 = 0.9$. The test statistic in the original method proposed by Fava et al. (2003), assigning equal weight to the two treatment differences, yielding a p-value of 0.168.

Score Tests: The values of the test statistics for score tests: for 1 DOF score test with $r = 1$, 2, and 5 and for the 2 DOF test, the values are 1.74, 2.52, 2.84, and 2.86, with corresponding p-values 0.19, 0.11, 0.09, and 0.24. The estimated response rates were $\hat{p_1} = 10/54 = 0.185$, $\hat{q_1} = 29/167 = 0.174$, $\hat{p_2} = 11/61 = 0.180$, $\hat{q_2} = 5/63 = 0.079$. The large estimated ρ, $\hat{\rho} = 8.8$ (which is probably considerably larger than will typically be observed in SPCD trials and probably resulted from the low dose used in the trial, and the fact that aripiprazole was adjunctive to other therapy), explains why the 1 DOF test with $r = 5$ yielded a smaller p-value than tests with $r =$ 1 and 2. The 2 DOF test is robust to r misspecification, when the true rates are equal to the observed ADAPT-A rates the 2 DOF test has better power than the 1 DOF test with r lower than 2 but worse than 1 DOF test with r higher than 2.

7.4.2 Continuous Response

SUR: The estimate of the treatment effect using SUR are obtained in SAS using PROC MODEL. The significant difference in the adjusted mean outcome is found for Phase II, while the difference in the adjusted mean outcome for Phase I is non-significant (Table 7.4). The treatment effect, defined as the weighted adjusted mean difference for the two Phases with weight $w = 0.75$ is not significant and estimated with 95% Confidence Interval to be -0.85 (-2.80, 1.09).

Table 7.4 ADAPT-A Estimates based on the SUR Model

	Parameter	Estimate	StdErr	tValue	Probt
Phase I	Intercept	−3.284	3.714	−0.880	0.378
	Baseline	−0.160	0.117	−1.360	0.176
	Treatment	−0.286	1.262	−0.230	0.821
Phase II	Intercept	−0.460	2.911	−0.160	0.875
	Baseline	−0.110	0.107	−1.030	0.303
	Treatment	−2.553	1.192	−2.140	0.034

Table 7.5 ADAPT-a estimates based on the OLS model

	Parameter	Estimate	StdErr	tValue	Probt
Phase I	Intercept	−3.223	3.715	−0.870	0.387
	Baseline	−0.161	0.117	−1.380	0.171
	Treatment	−0.289	1.262	−0.230	0.819
Phase II	Intercept	−0.813	2.912	−0.280	0.781
	Baseline	0.095	0.107	−0.890	0.376
	Treatment	−2.524	1.192	−2.120	0.036

OLS: The estimates of the treatment effect using OLS are obtained using two ANCOVA models fit in SAS PROC REG. Then the results are combined as above. As with SUR, no significant difference in the adjusted mean outcome is found for Phase I, while the difference in the adjusted mean outcome for Phase II is significant (Table 7.5). The combined treatment effect is not significant and estimated with 95% Confidence Interval to be −0.85 (−2.79, 1.10).

RMM: The estimates of the coefficients using RMM are obtained in SAS using PROC MIXED. As with the previous two methods, no significant difference in the adjusted mean outcome is found for Phase I, while the difference in the adjusted mean outcome for Phase II is significant (Table 7.6). There is no significant difference in the combined adjusted mean outcome. The combined treatment effect is estimated with 95% Confidence Interval to be −0.82 (−2.78, 1.13).

WRMM: The individual weights for Phase II were determined for Phase I placebo subjects via placebo response prediction model, which is based on the baseline MADRS score, and via K-means clustering, which is based on percent change from baseline in Phase I and Phase II baseline value. The Fig. 7.3 shows the weight characteristics for propensity-based approach. Panel 1 of the figure shows the weight change with the baseline measure (a slight jitter was added in order to show all subjects). Some placebo responders were assigned relatively high weights and some placebo non-responders were assigned low weights.

The next three panels of the figure show relationship of the weight and the baseline MADRS measure with main characteristics determining placebo response, that is, the percent change from baseline and the MADRS score at the end of Phase I. The

Table 7.6 ADAPT-a estimates based on the RMM

	Parameter	Estimate	StdErr	tValue	Probt
Phase I	Intercept	−3.233	3.714	−0.870	0.385
	Baseline	−0.161	0.117	−1.370	0.171
	Treatment	−0.268	1.262	−0.210	0.832
Phase II	Intercept	−1.223	2.871	−0.430	0.670
	Baseline	−0.077	0.105	−0.730	0.466
	Treatment	−2.491	1.175	−2.120	0.035
Phase II in placebo responders & Phase I drug	Intercept	−0.247	1.729	−0.140	0.887
	Baseline	−0.088	0.085	−1.040	0.299
	Treatment	−1.667	2.027	−0.820	0.412

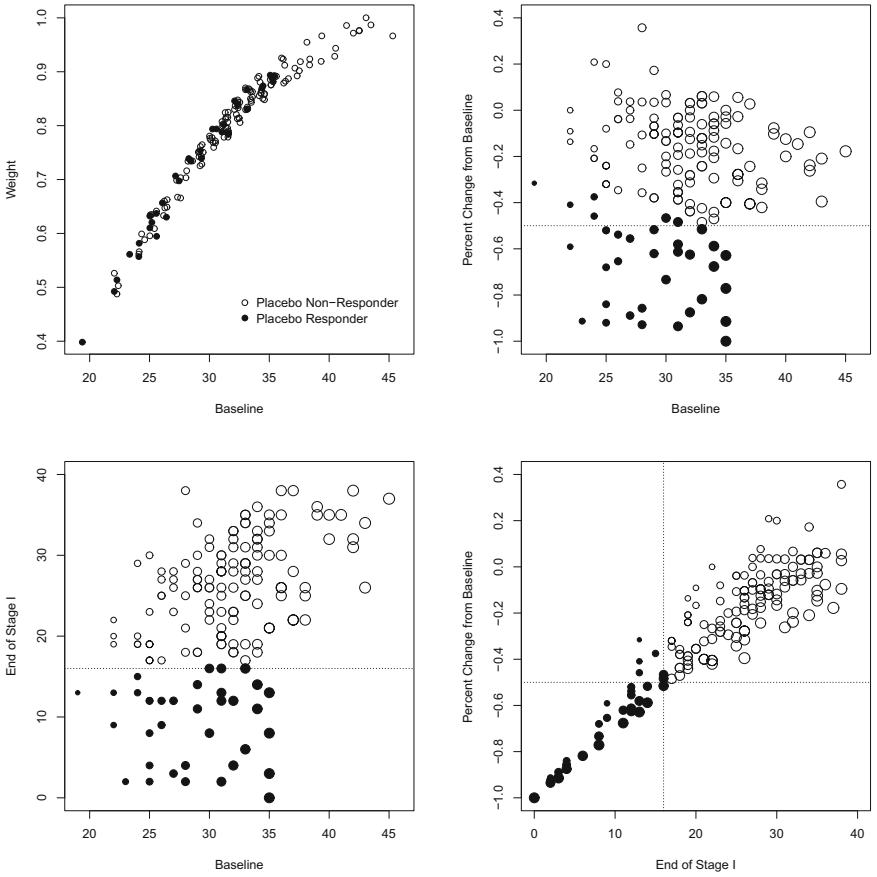


Fig. 7.3 ADAPT-a trial propensity-based weights for placebo responders and non-responders

vertical and horizontal lines represent the defined response criteria (50% decrease and 16 points MADRS score). The diameter of the dots is proportional to the weight. Panels 2 and 3 show that those with high baseline scores - namely above 35 points - tended to have smaller change in Phase I when randomized to placebo. Hence, those with more severe disease tended to have less placebo response.

Similarly, Fig. 7.4 presents weights based on K-means clustering ($K = 2, k = 1.5$). The weights assigned to placebo responders are visibly lower. This results in lower effective sample size and loss of power.

Table 7.7 presents a comparison of the estimates of the combined treatment effect among the four different methods: seemingly unrelated regression, ordinary least squares estimation, repeated measures model, and weighted repeated measures model using weighting of the subjects based on propensity to placebo response and K-means clustering. The difference is quite trivial in this case.

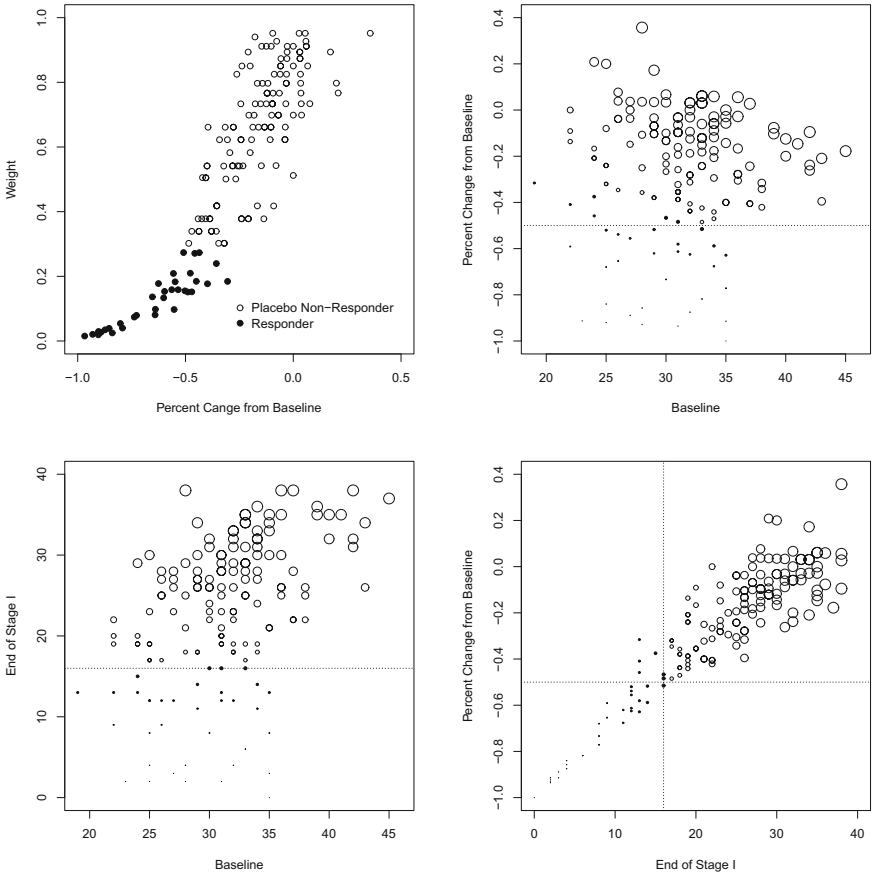


Fig. 7.4 ADAPT-A trial K-means based weights for placebo responders and non-responders

Table 7.7 ADAPT-a estimates of treatment effect based on the four methods

Method	Estimate	Statistic	P-Value
SUR	−0.853	−0.860	0.391
OLS	−0.848	−0.854	0.393
RMM	−0.824	−0.830	0.407
WRMM Propensity	−0.819	−0.830	0.407
WRMM CDF k = 0.5	−0.846	−0.850	0.394
WRMM CDF k = 1.0	−0.867	−0.880	0.382
WRMM CDF k = 1.5	−0.872	−0.880	0.380

7.5 Discussion

A typical placebo-controlled clinical trial in depressed patients might be eight weeks in duration whereas a sequential parallel design might be designed for 12 weeks. Thus, the trade-off involves more visits for the novel design versus more randomized patients in the conventional design. In depression trials conducted in practice, it is typical to space visit intervals more frequently early in the trial. As an example, for an eight week trial, visits might be scheduled after 1, 2, 4, 6, and 8 weeks. In a sequential parallel comparison design, it is believed that it is preferable to schedule visits at a constant visit interval such as after every two weeks and to blind the initiation of the second phase of the study to both investigators and participating subjects.

The gain in efficiency in this design is related to three factors, (1) the treatment effects, δ_i in each phase, (2) the variance covariance matrix Σ, and (3) the sample size of the second phase relative to the first phase of the study. The four approaches mentioned in this chapter provide different angles to view the data collected in the SPCD trials. The three factors are discussed extensively under various assumptions, which provide multiple choices to statisticians and physicians during the phase of trial design, for different situations in the clinical studies.

However, as pointed by Chen et al. (2011), it is also recognized that there are some practical challenges that are worth addressing cooperatively with clinicians and other clinical trial staffs in the future. First, although it is beneficial to re-randomize placebo non-responders into Phase II from a statistical point of view, such a re-randomization may threaten the integrity of the trial conduct. It may not be easy to maintain the blindness over the entire course of Phase II: it is likely that placebo non-responders who have taken placebo for the whole period of Phase I may feel the treatment effect right after they switch to the active treatment in Phase II if there is an influence due to the active treatment use; while those who continue on placebo may find their symptoms getting worse over time as they are placebo non-responders. The investigators may sense the treatments assigned to some subjects in Phase II if their symptoms appear to improve. Second, SPCD trials may take longer to complete than conventional trial, so its gain in power could be limited when the treatment difference is already very large at the end of Phase I. Finally, in addition to the practical issues

discussed above related to a SPCD design, there is a critically important issue as discussed in Chi et al. (2016) regarding the proper definition of the true treatment effect in a population Ω with a substantial proportion (α_R) of placebo responders Ω_R. If one also denotes the subpopulation of placebo non-responders by Ω_{NR} and its proportion by α_{NR}, then the relative treatment difference $\Delta_1 = \mu_{1,T} - \mu_{1,P}$ between treatment and placebo at the end of Phase I can also be represented as $\Delta_1 = \alpha_R \Delta_R + \alpha_{NR} \Delta_{NR}$, where $\alpha_{NR} = (1 - \alpha_R)$ and α_R is the proportion of placebo responders, $\Delta_R = \mu_{R,T} - \mu_{R,P}$ is the treatment effect between treatment and placebo among the placebo responders and $\Delta_{NR} = \mu_{NR,T} - \mu_{NR,P}$ is the treatment effect between treatment and placebo among the placebo non-responders. Therefore, when there is a substantial presence of placebo responders, then the term $\alpha_R \Delta_R$ is expected to be small and the relative treatment difference Δ_1 will under represent the true treatment effect. This is the primary reason why many previous trials in populations with a high placebo response rate using the traditional randomized parallel design had failed. On the other hand, with a SPCD design, sponsors may find it difficult to determine the weights on Phase I data (ω) and Phase II data (1 - ω) at the design stage. Since the original motivation for a SPCD design is based on the intuitive expectation that the treatment effect size at the end of Phase II is greater than that at the end of Phase I, it may be tempting to take advantage of the expectedly larger effect size at the end of Phase II and give a larger weight to Phase II data and thus biasing the result in favor of the treatment. For example, even with the Phase I randomization ratio $r_1 = 2$ as often used in recent SPCD trials, overweighting of Phase II data can occur. In addition, when the number of evaluable subjects in Phase II is small, putting too much weight on Phase II data may jeopardize the validity of the asymptotically normal inference from Phase II data and the interpretation and acceptability of the trial results. Thus, the choice of the weight ω is critically important in order to avoid the potential bias in overestimating the true treatment effect which may lead to subsequent approval of an ineffective treatment and if approved, with an incorrect dosing recommendation. Chi et al. (2016) propose to define the true treatment effect in a SPCD design as a weighted average of the treatment effects from Phase I and Phase II of a SPCD design using weights defined through the inverse variances following the method of weighted least square. This weighted average can be shown to actually represent the true treatment effect under certain reasonable and mild assumptions. An abbreviated discussion of their method is discussed in the next paragraph.

With the development of new methods in SPCD, there are also some discussions in the analysis methods. Ishida (2016) works on the OLS method for continuous responses. He examined the consequence of the Equal Covariance Assumption, which assumes the same covariance between Stage II baseline and endpoint for subjects in Placebo-Placebo and Placebo-Drug groups. With this assumption, if ANCOVA models are applied on an SPCD trial, then the estimates of the treatment effects are asymptotically uncorrelated. However, Ishida shows that if the Equal Covariance Assumption is not met, the proposed estimator statistic for the weighted treatment effect is biased. Li et al. (2016) is proposing an unbiased estimator of the Two-Period Treatment Effect, that is, an estimator for the difference in Drug-Drug and Placebo-Placebo group over Stage I and Stage II. They assume normal responses

in Stage I and Stage II, and that a non-responder is a subject with a good outcome: $Y_2 > C$. Then, they develop a moment estimator based on the proposed model. However, the correlation between the Stage I and Stage II data is not incorporated in the evaluation. With the proposed estimator, the adjustment for baseline or important covariates is not straightforward. In Chi et al.'s work (Chi et al. 2016), they pointed out a population Ω with a substantial proportion (α_R) of placebo responders Ω_R, one is often unable to characterize the subpopulation Ω_R. On the other hand, if the relative treatment effect in a randomized parallel design tends to underestimate the true treatment effect, then how is it possible to estimate the true treatment effect in an unbiased manner? Or put it in another way, how should one define the true treatment effect to be estimated when the relative treatment effect only represents a reduced true treatment effect? For this reason, Chi et al. (2016) call the relative treatment effect in a randomized parallel design an *apparent treatment effect*. They then introduce the concept of an *adjusted treatment effect* which is derived by adjusting the apparent treatment effect Δ_1 from Phase I of a SPCD design with Phase I randomization ratio $r_1 = 1$ by a quantity which is determined from the information from Phase II. Thus, the real value of a SPCD design lies in the fact that it provides additional information that will be needed for making the necessary and appropriate adjustment to the apparent treatment effect. This additional information is unavailable from a randomized parallel design as noted earlier, since one is unable to characterize the placebo responder subpopulation Ω_R. So, how is this additional information from Phase II being used to make the adjustment? Specifically, unlike in Li et al.'s work (Li et al. 2016), they assume that a placebo non-responder in Phase I is a subject with a response $X_{1,P}$ that is below certain threshold C. The true treatment effect Δ is then defined as a least squares weighted average of the treatment effect Δ_1 in Phase I with weight ω_1 and the treatment effect $(\Delta_2 \mid X_{1,P} < C)$ in Phase II with weight $\omega_2 = (1 - \omega_1)$, i.e., $\Delta = \omega_1\Delta_1 + \omega_2(\Delta_2 \mid X_{1,P} < C)$, under a SPCD design with a randomization ratio $r_1 = 1$ (i.e., with an equal allocation) in Phase I. The least square weight ω_2 as a function of r_1 is at its minimum when $r_1 = 1$ and does not overweight Phase II which avoids a potential common source of bias. This weight ω_2 is given by the expression, $\omega_2 = \left[1 + \left(\frac{\sigma_2}{\sigma_1}\right)^2 \frac{2}{\gamma}\right]^{-1}$, where ${\sigma_1}^2 = var(\hat{\Delta}_1), {\sigma_2}^2 = var(\hat{\Delta}_2 \mid X_{1,P} < C)$, and $\gamma = \Phi(\tau) = \Phi\left(\frac{C-\mu_{1,P}}{\sigma_{1,P}}\right)$ is the proportion of placebo responds in the population Ω, and where one simplifies the expression by taking advantage of the fact that $cov\left(\hat{\Delta}_1, (\hat{\Delta}_2 \mid X_{1,P} < C)\right) \to 0$ asymptotically. *It is important to point out here that the SPCD design with a randomization ratio* $r_1 = 1$ *is only used in defining the weight* ω_2, *while the actual SPCD design implemented can adopt a randomization ratio* $r_1 > 1$. This will not affect the definition of the weight ω_1, but in fact will allow a flexibility that can provide greater precision in the estimate of the treatment effect $(\Delta_2 \mid X_{1,P} < C)$ in Phase II and strengthen the validity of the statistical inference. They further show that the true treatment effect Δ represents the apparent treatment effect Δ_1 adjusted appropriately for the presence of placebo responders in the population as follows: In the derivation below, the following relationships are used, $\Delta_1 = \alpha_R\delta_R + \alpha_{NR}\Delta_{NR}$, $(\Delta_2 \mid X_{1,P} < C) = \Delta_{NR}$, $\omega_1 = (1 - \omega_2)$, and through cancel-

lations and simplification, one obtains the following expression for the true treatment effect, $\Delta = \omega_1\Delta_1 + \omega_2(\Delta_2 \mid X_{1,P} < C) = \omega_1[\alpha_R\Delta_R + \alpha_{NR}\Delta_{NR}] + \omega_2(\Delta_2 \mid X_{1,P} < C) \approx \omega_1[\alpha_R\Delta_R + \alpha_{NR}\Delta_{NR}] + \omega_2\Delta_{NR} = \Delta_1 + \omega_2\alpha_R(\Delta_{NR} - \Delta_R)$. Now from the above equation, $\Delta = \Delta_1 + \omega_2[\alpha_R(\Delta_{NR} - \Delta_R)]$, one can make a few observations. If there are no placebo responders, then $\alpha_R = 0$, then $\Delta = \Delta_1$, that is, the adjusted treatment effect Δ and the apparent treatment effect Δ_1 are identical and hence no adjustment is needed. On the other hand, if Ω_R is not empty, then it is expected that $\Delta_{NR} > \Delta_R$. In this case, then the expression, $[\alpha_R(\Delta_{NR} - \Delta_R)]$, represents the total amount of expected treatment effect Δ_{NR} that is not observed due to the placebo response in Ω_R. Now because $\Delta_{NR} \approx (\Delta_2 \mid X_{1,P} < C)$, one can view $[\alpha_R(\Delta_{NR} - \Delta_R)] = [\alpha_R((\Delta_2 \mid X_{1,P} < C) - \Delta_R)]$ as the equivalent amount of treatment effect from Phase II that has been nullified by the placebo response in Ω_R. Then it follows that $\omega_2[\alpha_R(\Delta_{NR} - \Delta_R)]$ represents the appropriately weighted amount of the treatment effect, $[\alpha_R((\Delta_2 \mid X_{1,P} < C) - \Delta_R)]$, from Phase II that needs to be added to the apparent treatment effect Δ_1 from Phase I to account for the presence of placebo responders Ω_R. Hence the amount of adjustment, $\omega_2[\alpha_R(\Delta_{NR} - \Delta_R)]$, properly compensates for the presence of placebo responders. Furthermore, in order for the adjusted treatment effectiveness claim to be extendable to the intended study population, the authors introduce a consistency measure for assessing the consistency between the treatment effect Δ_1 from Phase I and the treatment effect $(\Delta_2 \mid X_{1,P} < C)$ from Phase II. They propose to jointly test the consistency and the efficacy hypothesis. In light of the proposed adjustment and the joint test of efficacy and consistency, the sample size requirement may be higher than one may wish. But that is the cost one has to pay for the additional information needed from Phase II of a SPCD design to make the necessary adjustment for the presence of placebo responders and to allow the efficacy claim to be extendable to the intended population. For a detailed discussion of the various issues associated with a SPCD design and proposed methodology, the interested readers may refer to their original paper in Chi et al. (2016).

References

A paroxetine- and placebo-controlled study of 50 mg/day and 100 mg/day of EB-1010 among outpatients with major depressive disorder who have responded inadequately to prior Selective Serotonin Reuptake Inhibitors (SSRIs) and Serotonin Norepinephrine Reuptake Inhibitors (SNRIs). https://clinicaltrials.gov/show/NCT01318434. Accessed: July 18, 2014.

Alkermes announces advances with its late-stage CNS pipeline. http://phx.corporate-ir.net/phoenix.zhtml?c=92211&p=irol-newsArticlePrint&ID=1888936&highlight. Accessed Augest 01, 2014.

Bridge, J. A., Birmaher, B., Iyengar, S., Barbe, R. P., & Brent, D. A. (2009). Placebo response in randomized controlled trials of antidepressants for pediatric major depressive disorder. *American Journal of Psychiatry*, *166*(1), 42–49.

Carmody, T. J., Rush, A. J., Bernstein, I., Warden, D., Brannan, S., Burnham, D., et al. (2006). The Montgomery Äsberg and the Hamilton ratings of depression: a comparison of measures. *European Neuropsychopharmacology*, *16*(8), 601–611.

Chen, J. A., Vijapura, S., Papakostas, G. I., Parkin, S. R., Hyung, D. J., Kim, C. C., et al. (2015). Association between physician beliefs regarding assigned treatment and clinical response: Re-analysis of data from the hypericum depression trial study group. *Asian Journal of Psychiatry*, *13*, 23–29.

Chen, Y.-F., Yang, Y., Hung, H. M. J., & Wang, S.-J. (2011). Evaluation of performance of some enrichment designs dealing with high placebo responce in psychiatric clinical trials. *Contemporary Clinical Trials*, *32*(4), 592–604.

Chen, Y.-F., Zhang, X., Tamura, R. N., & Chen, C. M. (2014). A sequential enriched design for target patient population in psychiatric clinical trials. *Statistics in Medicine*, *33*(17), 2953–2967.

Chi, G., Li, Y., Liu, Y., Lewin, D., & Lim, P. (2016). On clinical trials with a high placebo response rate. *Contemporary Clinical Trials Communications*, *2*, 34–53.

Cressey, D. (2011). Psychopharmacology in crisis. Nature, 2011-10.

Doros, G., Pencina, M., Rybin, D., Meisne, A., & Fava, M. (2013). A repeated measures model for analysis of continuous outcomes in sequential parallel comparison design studies. *Statistics in Medicine*, *32*(16), 2767–2789.

Douglas, E. F., Heiligenstein, J. H., Tollefson, G. D., & Potter, W. Z. (2001). The double-blind variable placebo lead-in period: Results from two antidepressant clinical trials. *Journal of Clinical Psychopharmacology*, *21*(6), 561–568.

Fava, M., Mischoulon, D., Iosifescu, D., Witte, J., Pencina, M., Flynn, M., Harper, L., Levy, M., Rickels, K., & Pollack, M. (2009). A double-blind, placebo-controlled study of aripiprazole adjunctive to antidepressant therapy (ADT) among depressed outpatients with inadequate response to prior ADT (ADAPT-A study). In The 48th Annual Meeting of the American College of Neuropsychopharmacology.

Fava, M., Evins, A. E., Tollefson, G. D., & Potter, W. Z. (2003). The problem of the placebo response in clinical trials for psychiatric disorders: Culprits, possible remedies, and a novel study design approach. *Psychotherapy and Psychosomatics*, *72*, 115–127.

Fava, M., Mischoulon, D., Iosifescu, D., Witte, J., Pencina, M., Flynn, M., et al. (2012). A double-blind, placebo-controlled study of aripiprazole adjunctive to antidepressant therapy among depressed outpatients with inadequate response to prior antidepressant therapy (ADAPT-A study). *Psychotherapy and Psychosomatics*, *81*, 87–97.

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., et al. (2010). Antidepressant drug effects and depression severity: A patient-level meta-analysis. *JAMA*, *303*(1), 47–53.

Huang, X., & Tamura, R. N. (2010). Comparison of test statistics for the sequential parallel design. *Statistics in Biopharmaceutical Research*, *2*(1), 42–50.

Ishida, E. (2016). *Placebo response and sequential parallel comparison design (SPCD)*. Atlanta, GA: In International Chineses Statistical Assiciation.

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

Khan, A., Detke, M., Khan, S. R. F., & Mallinckrodt., C. (2003). Placebo response and antidepressant clinical trial outcome. *The Journal of Nervous and Mental Disease*, *191*(4), 211–218.

Khan, A., Khan, S. R., Shankles, E. B., & Polissar, N. L. (2002). Relative sensitivity of the Montgomery-Äsberg depression rating scale, the Hamilton depression rating scale and the clinical global impressions rating scale in antidepressant clinical trials. *International Clinical Psychopharmacology*, *17*(6), 281–285.

Khan, A., Khan, S. R., Walens, G., Kolts, R., & Giller, E. L. (2003). Frequency of positive studies among fixed and flexible dose antidepressant clinical trials: An analysis of the food and drug administraton summary basis of approval reports. *Neuropsychopharmacology*, *28*(3), 552.

Khan, Arif, Kolts, R. L., Thase, M. E., Krishnan, K. R. R., & Brown, W. (2004). Research design features and patient characteristics associated with the outcome of antidepressant clinical trials. *American Journal of Psychiatry*, *161*(11), 2045–2049.

Khin, N. A., Chen, Y.-F., Yang, Y., Yang, P., & Laughren, T. P. (2011). Exploratory analyses of efficacy data from major depressive disorder trials submitted to the US Food and Drug Administration in support of new drug applications. *The Journal of Clinical Psychiatry*, *72*(4), 464–472.

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the Food and Drug Administration. *PLoS medicine*, *5*(2), e45.

Krell, H.V., Leuchter, A.F. Morgan, M., Cook, I.A., & Abrams, M. (2004). Subject expectations of treatment effectiveness and outcome of treatment with an experimental antidepressant. *The Journal of Clinical Psychiatry*.

Landin, R., DeBrota, D. J., DeVries, T. A., Potter, W. Z., & Demitrack, M. A. (2000). The impact of restrictive entry criterion during the placebo lead-in period. *Biometrics*, *56*(1), 271–278.

Li, Y., Liu, Y., Liu, Q., & Lim, P. (2016). *An unbiased estimator of the two-period treatment effect in doubly randomized delayed-start (DRDA) designs*. Atlanta, GA: In International Chineses Statistical Assiciation.

Liu, Q., Lim, P., Singh, J., Lewin, D., Schwab, Barry, & Kent, J. (2012). Doubly randomized delayed-start design for enrichment studies with responders or nonresponders. *Journal of Biopharmaceutical Statistics*, *22*(4), 737–757.

Meyer, B., Pilkonis, P. A., Krupnick, J. L., Egan, M. K., Simmens, S. J., & Sotsky, S. M. (2002). Treatment expectancies, patient alliance and outcome: Further analyses from the National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology*, *70*(4), 1051.

Mundt, J. C., Greist, J. H., Jefferson, J. W., Katzelnick, D. J., DeBrota, D. J., Chappell, P. B., et al. (2007). Is it easier to find what you are looking for if you think you know what it looks like? *Journal of Clinical Psychopharmacology*, *27*(2), 121–125.

Nutt, D., & Goodwin, G. (2011). ECNP Summit on the future of CNS drug research in Europe 2011: Report prepared for ECNP by David Nutt and Guy Goodwin. *European Neuropsychopharmacology*, *21*(7), 495–499.

Papakostas, G. I., & Fava, M. (2009). Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind, randomized clinical trials in MDD. *European Neuropsychopharmacology*, *19*(1), 34–40.

Papakostas, G. I., Shelton, R. C., Zajecka, J. M., Etemad, B., Rickels, K., Clain, A., et al. (2012). L-methylfolate as adjunctive therapy for SSRI-resistant major depression: Results of two randomized, double-blind, parallel-sequential trials. *American Journal of Psychiatry*, *169*(12), 1267–1274.

Posternak, M. A., & Zimmerman, M. (2007). Therapeutic effect of follow-up assessments on antidepressant and placebo response rates in antidepressant efficacy trials. *The British Journal of Psychiatry*, *190*(4), 287–292.

Rief, W., Nestoriuc, Y., Weiss, S., Welzel, E., Barsky, A. J., & Hofmann, S. G. (2009). Meta-analysis of the placebo response in antidepressant trials. *Journal of Affective Disorders*, *118*(1), 1–8.

Robinson, D. S., & Rickels, K. (2000). Concerns about clinical drug trials. *Journal of Clinical Psychopharmacology*, *20*(6), 593–596.

Rutherford, B. R., Marcus, S. M., Wang, P., Sneed, J. R., Pelton, G., Devanand, D., et al. (2013). A randomized, prospective pilot study of patient expectancy and antidepressant outcome. *Psychological Medicine*, *43*(5), 975–982.

Rutherford, B. R., & Roose, S. P. (2013). A model of placebo response in antidepressant clinical trials. *American Journal of Psychiatry*, *170*(7), 723–733.

Rutherford, B. R., Sneed, J. R., & Roose, S. P. (2009). Does study design influence outcome? *Psychotherapy and Psychosomatics*, *78*(3), 172–181.

Rutherford, B. R., Sneed, J. R., Tandler, J. M., Rindskopf, D., Peterson, B. S., & Roose, S. P. (2011). Deconstructing pediatric depression trials: an analysis of the effects of expectancy and therapeutic contact. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50*(8), 782–795.

Rybin, D., Doros, G., Pencina, M. J., & Fava, M. (2015). Placebo non-response measure in sequential parallel comparison design studies. *Statistics in Medicine*, *34*(15), 2281–2293.

Sneed, J. R., Rutherford, B. R., Rindskopf, D., Lane, D. T., Sackeim, H. A., & Roose, S. P. (2008). Design makes a difference: A meta-analysis of antidepressant response rates in placebo-controlled versus comparator trials in late-life depression. *The American Journal of Geriatric Psychiatry*, *16*(1), 65–73.

Sotsky, S. M., Glass, D. R., Tracie Shea, M., Pilkonis, P. A., Collins, J. F., Elkin, I., et al. (1991). Patient predictors of response to psychotherapy and pharmacotherapy: Findings in the NIMH Treatment of Depression Collaborative Research Program. *The American Journal of Psychiatry*, *148*(8), 997–1008.

Stein, D. J., Baldwin, D. S., Dolberg, O. T., Despiegel, N., & Bandelow, B. (2006). Which factors predict placebo response in anxiety disorders and major depression? An analysis of placebo-controlled studies of escitalopram. *The Journal of Clinical Psychiatry*, *67*(11), 1741–1746.

Study of 6(S)-5-MTHF among selective serotonin reuptake inhibitorresistant outpatients with major depressive disorder (TRD-2). http://clinicaltrials.gov/show/NCT00955955. Accessed: July 18, 2014.

Tamura, R. N., & Huang, X. (2007). An examination of the efficiency of the sequential parallel design in psychiatric clinical trials. *Clinical Trials*, *4*, 309–317.

Tamura, R. N., Huang, X., & Boos, D. D. (2011). Estimation of treatment effect for the sequential parallel design. *Statistics in Medicine*, *30*(30), 3496–3506.

Thase, M. E., (1999). How should efficacy be evaluated in randomized clinical trials of treatments for depression? In *Assessing Antidepressant Efficacy: A Reexamination., Jan 1998* (p. 1999). Phoenix, AZ, US: Physicians Postgraduate Press.

Trivedi, M. H., & Rush, J. (1994). Does a placebo run-in or a placebo treatment cell affect the efficacy of antidepressant medications? *Neuropsy- chopharmacology*, *11*(1), 33–43.

Wernicke, J. F., Sayler, M. E., Koke, S. C., Pearson, D. K., & Tollefson, G. D. (1997). Fluoxetine and concomitant centrally acting medication use during clinical trials of depression: The absence of an effect related to agitation and suicidal behavior. *Depression and Anxiety*, *6*(1), 31–39.

Wied, C. G., Stoyanova, V., Yu, Y., Isaac, M., Pani, L., & de Andres-Trelles, F. (2012). The placebo arm in clinical studies for treatment of psychiatric disorders: A regulatory dilemma. *European Neuropsychopharmacology*, *22*(11), 804–811.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, *57*(298), 348–368.

Chapter 8
Phase I Cancer Clinical Trial Design: Single and Combination Agents

Ying Yuan, Heng Zhou and Yanhong Zhou

8.1 Introduction

The objective of a phase I trial is to find the maximum tolerated dose (MTD), which is defined as the dose or dose combination with the toxicity probability closest to the target toxicity rate. A phase I clinical trial is critically important because it determines the MTD that will be further investigated in the subsequent phase II or III trials. Misidentification of the MTD could result in an inconclusive trial, thereby wasting enormous resources, or lead to a trial in which a substantial number of patients are treated at excessively toxic doses. In addition, inappropriate selection of a dose as the MTD might cause researchers to overlook a promising drug if the dose has low toxicity and negligible efficacy.

Numerous phase I trial designs have been proposed to find the MTD. These designs are generally classified into algorithm-based designs and model-based designs (Jaki et al. 2013; van Brummelen et al. 2016). The algorithm-based design uses a set of simple, pre-specified rules (or algorithm) to determine dose escalation and de-escalation, without assuming any model on the dose-toxicity curve. Examples include the 3+3 design (Storer 1989, 2001) and up-and-down design (Stylianou and Flournoy 2002). The major advantages of the algorithm-based design, such as the 3+3 design, are transparency and simplicity. The implementation of the design does not require a computer program or much support from statisticians. Despite widespread criticism of the 3+3 design for poor operating characteristics, its simplicity continues to make it the dominant phase I trial design used in practice.

Y. Yuan (✉) · H. Zhou · Y. Zhou
Department of Biostatistics, The University of Texas MD Anderson Cancer Center,
Houston, TX 77030, USA
e-mail: yyuan@mdanderson.org

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_8

Model-based dose-finding designs have been proposed in order to achieve a better performance of dose finding. Model-based designs often assume a parametric dose-toxicity model indexed by one or two parameters, such as a probability power function or a logistic regression model. As information accrues during the trial, the dose-toxicity relationship is re-evaluated by updating the estimates of the model parameters and then used to guide the dose allocation for subsequent patients. A typical example of the model-based design is the continuous reassessment method (CRM) (Pepe et al. 1990). Although a model-based design, such as the CRM, yields better performance than an algorithm-based design (Jaki et al. 2013; van Brummelen et al. 2016; Iasonos and O'Quigley 2014), it is considered by many to be statistically and computationally complex and its implementation requires repeated model fitting and estimation. This leads practitioners to perceive dose allocations as coming from a "black box". As a result, the use of the model-based designs has been fairly limited in practice (Rogatko et al. 2007).

There has been increasing interest in a new class of designs that combine the simplicity of algorithm-based designs and the good performance of model-based designs. This class of designs utilizes a model to derive the design, similar to the model-based design, but its rule of dose escalation and de-escalation can be pre-tabulated before the onset of the trial in a fashion similar to the algorithm-based design. We refer to this new class of designs as the "model-assisted" design. Examples of model-assisted designs include the modified toxicity probability interval (mTPI) design (Ji et al. 2010), Bayesian optimal interval (BOIN) design (Liu and Yuan 2015; Yuan et al. 2016), and keyboard design (Yan et al. 2017). In what follows, we will first focus on single-agent phase I trials and then describe drug combination trials.

8.2 Single-Agent Trials

8.2.1 Algorithm-Based Designs

The most widely used algorithm-based design is the 3+3 design. Actually, the 3+3 design is a family of designs, for which there are numerous variations. Table 8.1 gives a specification of a 3+3 design that is commonly used as a boilerplate in many phase I protocols. Although the algorithm presented in Table 8.1 is often included in a protocol to describe the 3+3 design, it is actually incomplete. For example, it does not say what to do or what to conclude if either 2 out of 3 dose-limiting toxicities (DLTs) are seen at the lowest dose level or 0 out of 3 DLT is observed at the highest dose level; see Yuan et al. (2016) for details. The 3+3 design has been widely criticized for its poor operating characteristics, e.g., poor accuracy in identifying the MTD and a great tendency to underdose patients. For this reason, we do not discuss the 3+3 design further.

Table 8.1 A common boilerplate phase I protocol 3+3 algorithm

Number of patients with DLT at a given dose level	Escalation decision rule
0/3	Enter 3 patients at the next dose level
1/3	Enter at least 3 more patients at this dose level
	• If 0 of these 3 patients experiences DLT, proceed to the next dose level
	• If ≥ 1 of this group suffers DLT, this dose exceeds the MTD and dose escalation is stopped
	3 additional patients will be entered at the next lower dose level if only 3 patients were treated previously at that dose
≥ 2	Dose escalation will be stopped. This dose level will be declared the maximally administered dose (highest dose administered). Three (3) additional patients will be entered at the next lower dose level if only 3 patients were treated previously at that dose

MTD: The highest dose at which no more than 1 of 6 evaluable patients has had DLT. Six patients should be treated before the dose is declared as the MTD

8.2.2 Model-Based Designs

8.2.2.1 Continuous Reassessment Method (CRM)

The CRM is a typical example of a model-base design. Let $(d_1, \ldots, d_J)$ denote a set of J prespecified doses for the drug under investigation, and ϕ denote the target toxicity rate specified by physicians. The CRM assumes a parametric model for the dose-toxicity curve. Based on the accrued data, the CRM continuously updates the model estimate of the dose-toxicity curve and makes the decision of dose assignment for the new patients. Commonly used dose-toxicity models include the power model and logistic model. Specifically, the power model is given by

$$\text{pr}(\text{toxicity at } d_j) = p_j(\alpha) = \pi_j^{\exp(\alpha)} \qquad j = 1, \ldots, J,$$

where $(\pi_1, \ldots, \pi_J)$ are the prior estimates of the toxicity probabilities (often known as the "skeleton") at the J doses, and α is an unknown parameter. Research shows that the choice of the power model or logistic model has little impact on the performance of the CRM. What is more critical is the specification of each model, e.g., the specification of the skeleton.

Suppose that among n_j patients treated at dose level j, y_j patients have experienced DLT. Let D denote the observed data, $D = \{(n_j, y_j), j = 1, \ldots, J\}$. Based on the binomial distribution for the toxicity outcome, the likelihood function is given by

$$L(D|\alpha) = \prod_{j=1}^{J} \{\pi_j^{\exp(\alpha)}\}^{y_j} \{1 - \pi_j^{\exp(\alpha)}\}^{n_j - y_j}.$$

The DLT rate of dose d_j is estimated by its posterior mean

$$\hat{p}_j = \int \pi_j^{\exp(\alpha)} \frac{L(D|\alpha) f(\alpha)}{\int L(D|\alpha) f(\alpha) d\alpha} d\alpha,$$

where $f(\alpha)$ is a prior distribution for the parameter α, often assumed to follow a normal distribution $N(0, 2)$.

The dose-finding algorithm of the CRM is described below.

1. Patients in the first cohort are treated at the lowest dose d_1, or the physician-specified dose.
2. Based on the cumulated data, we obtain the posterior DLT rate estimate $\hat{p}_j$, and find dose level j^* that has a DLT rate closest to ϕ. Let j denote the current dose level. If $j^* < j$, we de-escalate the dose level to $j - 1$; if $j^* > j$, we escalate the dose level to $j + 1$; otherwise, the dose stays at the dose level j for the next cohort of patients.
3. Once the maximum sample size N is reached, the trial is completed and the dose that has the DLT rate closest to ϕ is selected as the MTD.

In practice, we often impose an early stopping rule: If pr(toxicity rate at $d_1 > \phi|D$) > 0.9, the trial is terminated for safety.

Numerous studies have shown that the CRM has substantially better performance than the 3+3 design, however, the use of the CRM remains limited for several reasons. Because of the statistical and computational complexity of the CRM, communicating to clinical investigators how the design works remains challenging, which leads them to perceive the dose allocations as coming from a "black box". From a methodological viewpoint, the model-based CRM, although generally robust, is still subject to the influence of model specification. To obtain good operating characteristics, the CRM model (e.g., the skeleton) must be appropriately calibrated, which is a challenging procedure that requires extensive statistical expertise. To simplify the model calibration, Lee and Cheung (2009) proposed a systematic method to generate a "default" skeleton for the CRM. Their method requires users to specify only a half width of an indifferent interval and the prior location of the MTD, and can be easily carried out using the "getprior()" function in the R package "dfcrm". Lee and Cheung's method simplifies the specification of the skeleton, but the issue of model sensitivity remains. Table 8.2 shows the simulation results of the CRM with two different skeletons, skeleton 1 = (0.070, 0.127, 0.200, 0.286, 0.377, 0.468) and skeleton 2 = (0.012, 0.069, 0.200, 0.380, 0.560, 0.706), generated by using the method of Lee and Cheung with a half-width indifferent interval of 0.04 and 0.08. We can see that skeleton 1 substantially outperforms skeleton 2 in scenario 1, whereas the opposite result is seen in scenario 2. In other words, a skeleton that works well in one scenario may not work as well in another scenario, and there does not exist a single "best" skeleton that dominates all others.

Table 8.2 Performance of the CRM and Bayesian model-averaging-CRM (BMA-CRM) with two different skeletons generated with half-width indifferent intervals of 0.04 and 0.08. The target toxicity rate $\phi = 0.2$ and sample size $N = 36$

Scenario 1							
True DLT rate		0.03	0.04	0.05	0.06	0.07	0.20
CRM with skeleton 1	% sel[a]	0	0.10	1.25	3.85	21.05	**73.20**
	# pts[b]	1.5	1.7	2.2	3.5	8.0	18.8
CRM with skeleton 2	% sel	0.05	1.35	5.70	8.10	28.25	**56.40**
	# pts	1.5	2.2	3.4	5.2	10.1	13.4
BMA-CRM	% sel	0	0.25	2.45	5.65	21.80	**69.60**
	# pts	1.5	1.9	2.7	3.9	8.6	17.4
Scenario 2							
True DLT rate		0.12	0.24	0.33	0.60	0.70	0.80
CRM with skeleton 1	% sel	35.15	**43.50**	10.45	0.25	0	0
	# pts	13.9	12.2	5.2	1.2	0.3	0.1
CRM with skeleton 2	% sel	30.60	**53.10**	11.15	0	0	0
	# pts	12.3	15.1	6.1	1.0	0.1	0
BMA-CRM	% sel	32.30	**49.85**	10.05	0	0	0
	# pts	12.7	14.3	5.5	1.0	0.3	0.1

[a]Average selection percentage at each dose
[b]Average number of patients treated at each dose

8.2.2.2 Bayesian Model-Averaging CRM (BMA-CRM)

One approach to reduce the sensitivity of the CRM to the skeleton is the BMA-CRM (Yin and Yuan 2009), which prespecifies multiple skeletons, each of which leads to a CRM model of the form (Sect. 8.2.2.1) with a different set of p_j's. The idea is to let the data determine which skeleton or model fits the data better and then automatically favor that model as the basis for making the decision of dose escalation and de-escalation.

Let $(M_1, \ldots, M_K)$ be the models that correspond to the K prespecified skeletons $\{(\pi_{11}, \ldots, \pi_{1J}), \ldots, (\pi_{K1}, \ldots, \pi_{KJ})\}$, where M_k $(k = 1, \ldots, K)$ takes a form

$$p_{kj}(\alpha_k) = \pi_{kj}^{\exp(\alpha_k)}, \quad j = 1, \ldots, J,$$

obtained using the kth skeleton $(\pi_{k1}, \ldots, \pi_{kJ})$. Let $\text{pr}(M_k)$ be the prior probability that model M_k is the true model, i.e., the probability that the kth skeleton $(\pi_{k1}, \ldots, \pi_{kJ})$ matches the true dose-toxicity curve. If there is no preference a priori for any single model in the CRM case, we can assign equal weights to the different skeletons by simply setting $\text{pr}(M_k) = 1/K$. When there is prior information about the importance of each set of the prespecified toxicity probabilities, we can incorporate such information into $\text{pr}(M_k)$. For example, if a certain set of the prespecification is more likely to be true, we can assign it a higher prior model probability.

At a certain stage of the trial, based on the observed data $D = \{(n_j, y_j), j = 1, \ldots, J\}$, the likelihood function under model M_k is

$$L(D|\alpha_k, M_k) = \prod_{j=1}^{J} \{\pi_{kj}^{\exp(\alpha_k)}\}^{y_j} \{1 - \pi_{kj}^{\exp(\alpha_k)}\}^{n_j - y_j}.$$

The posterior model probability for M_k is given by

$$\text{pr}(M_k|D) = \frac{L(D|M_k)\text{pr}(M_k)}{\sum_{i=1}^{K} L(D|M_i)\text{pr}(M_i)},$$

where $L(D|M_k)$ is the marginal likelihood of model M_k,

$$L(D|M_k) = \int L(D|\alpha_k, M_k) f(\alpha_k|M_k) d\alpha_k,$$

α_k is the power parameter in the CRM associated with model M_k, and $f(\alpha_k|M_k)$ is the prior distribution of α_k under model M_k.

The BMA estimate for the DLT rate at each dose level is given by

$$\bar{p}_j = \sum_{k=1}^{K} \hat{p}_{kj} \text{pr}(M_k|D), \quad j = 1, \ldots, J, \tag{8.1}$$

where $\hat{p}_{kj}$ is the posterior mean of the DLT rate of dose level j under model M_k, i.e.,

$$\hat{p}_{kj} = \int \pi_{kj}^{\exp(\alpha_k)} \frac{L(D|\alpha_k, M_k) f(\alpha_k|M_k)}{\int L(D|\alpha_k, M_k) f(\alpha_k|M_k) d\alpha_k} d\alpha_k.$$

By assigning $\hat{p}_{kj}$ a weight of $\text{pr}(M_k|D)$, the BMA method automatically identifies and favors the best fitting model, thus $\bar{p}_j$ is always close to the best estimate. The dose-finding algorithm of the BMA-CRM is the same as that of the CRM described previously, except that the DLT rate estimate $\hat{p}_j$ is replaced by the BMA estimate $\bar{p}_j$ given in (8.1). In other words, the decision of dose escalation or de-escalation in the trial is based upon $\bar{p}_j$ (which is determined through multiple skeletons) as opposed to $\hat{p}_j$ (which is based on a single skeleton).

Table 8.2 shows the performance of the BMA-CRM with two skeletons versus that of the CRM using one skeleton. The BMA-CRM is more reliable than the CRM, and its performance is close to that of the CRM with the better skeleton in both scenarios 1 and 2. Such robustness and reliability is important because in practice we often prefer a method that yields reliable performance to a method that has high variability (i.e., performs well in some scenarios, but not in other scenarios).

8.2.3 Model-Assisted Designs

Model-assisted designs comprise a new class of designs that combine the simplicity of algorithm-based designs and the good performance of model-based designs. This class of designs utilizes a statistical model (e.g., a binomial model) to derive the design, similar to the model-based design, but its rule of dose escalation and de-escalation can be pre-tabulated before the onset of the trial in a fashion similar to the algorithm-based design. This feature makes the model-based designs attractive in practice. In what follows, we introduce three different model-assisted designs, namely, the mTPI design, BOIN design, and keyboard design. The comparison of their performance is provided later in Sect. 8.2.4.

8.2.3.1 Modified Toxicity Probability Interval (mTPI) Design

The mTPI design starts by defining three dosing intervals: the underdosing interval $(0, \delta_1)$, proper dosing interval (δ_1, δ_2), and overdosing interval $(\delta_2, 1)$. For example, given the target DLT rate of 0.2, the three intervals may be respectively defined as (0, 0.15), (0.15, 0.25), and (0.25, 1) in terms of the DLT rate.

Suppose that at the current dose level j, y_j of n_j patients have experienced DLT. The mTPI design assumes that y_j follows a beta-binomial model:

$$\begin{aligned} y_j \mid n_j, p_j &\sim \text{Binom}(n_j, p_j) \\ p_j &\sim \text{Beta}(1, 1) \equiv \text{Unif}(0, 1). \end{aligned} \tag{8.2}$$

Then, the posterior distribution of p_j is given by

$$p_j \mid (y_j, n_j) \sim \text{Beta}(y_j + 1, n_j - y_j + 1), \text{ for } j = 1, \ldots, J. \tag{8.3}$$

The mTPI design makes the decision of dose escalation and de-escalation based on the unit probability mass (UPM) of the three intervals, defined as

$$\begin{aligned} \text{UPM}_{(0,\delta_1)} &= \text{pr}(p_j \in (0, \delta_1) \mid D_j)/\delta_1, \\ \text{UPM}_{(\delta_1,\delta_2)} &= \text{pr}(p_j \in (\delta_1, \delta_2) \mid D_j)/(\delta_2 - \delta_1), \\ \text{UPM}_{(\delta_2,1)} &= \text{pr}(p_j \in (\delta_2, 1) \mid D_j)/(1 - \delta_2). \end{aligned}$$

That is, given a specific dosing interval, the UPM is defined as the posterior probability of the interval divided by the length of the interval.

For treating the next patient, the dose escalation and de-escalation rule of the mTPI design is given below.

- Escalate the dose if $\text{UPM}_{(0,\delta_1)}$ (i.e., the UPM of the underdosing interval) is the largest,
- De-escalate the dose if $\text{UPM}_{(\delta_2,1)}$ the (i.e., the UPM of the overdosing interval) is the largest,
- Stay at the same dose if $\text{UPM}_{(\delta_1,\delta_2)}$ (i.e., the UPM of the proper dosing interval) is the largest.

One attractive feature of the mTPI design is that its decision rule can be pre-calculated for each possible $n_j = 1, \ldots, N$, which makes it easy to implement in practice.

The mTPI design stops the trial when the maximum sample size N is reached. The MTD is selected based on $\{\tilde{p}_j\}$, the isotonically transformed values of the observed DLT rates $\{\hat{p}_j\}$. Specifically, we select as the MTD dose j^*, for which the isotonic estimate of the DLT rate $\tilde{p}_{j^*}$ is closest to the target DLT rate ϕ. If there are ties for $\tilde{p}_{j^*}$, we select from the ties the highest dose level when $\tilde{p}_{j^*} < \phi$ or the lowest dose level when $\tilde{p}_{j^*} > \phi$. The isotonic estimates $\{\tilde{p}_j\}$ can be obtained by applying the pooled adjacent violators algorithm (Barlow et al. 1973) to $\{\hat{p}_j\}$. Operatively, the pooled adjacent violators algorithm replaces any adjacent $\hat{p}_j$'s that violate the non-decreasing order by their (weighted) average so that the resulting estimates $\tilde{p}_j$ become monotonic. In the case in which the observed DLT rates are monotonic, $\tilde{p}_j$ and $\hat{p}_j$ are equivalent.

One issue for the mTPI design is that it has a high risk of overdosing patients. For example, given the target DLT rate of 0.3, even when we observe that $3/6 = 0.5$ (50% of patients experienced DLTs), the mTPI design continues to treat the next cohort of patients at the same dose, which is intuitively excessively risky. The aggressiveness of the mTPI design stems from its use of the UPM as the criterion for determining dose allocation. To see this problem, consider a trial for which the target toxicity rate is 0.2, and the underdosing, proper dosing and overdosing intervals are (0, 0.17), (0.17, 0.23), and (0.23, 1), respectively. Suppose that at a certain stage of the trial, the observed data indicate that the posterior probabilities of the underdosing interval, proper dosing interval and overdosing interval are 0.01, 0.09 and 0.9, respectively. That is, there is 90% chance that the current dose is overdosing patients and only 9% chance that the current dose provides proper dosing. Despite such dominant evidence of overdosing, the mTPI design retains the same dose for treating the next new patient because the UPM for the proper dosing interval is the largest. Specifically, the UPM for the proper dosing interval is $0.09/(0.23 - 0.17) = 1.5$, and the UPM for the overdosing interval is $0.9/(1 - 0.23) = 1.17$. This example demonstrates that the UPM cannot appropriately quantify the evidence of overdosing.

8.2.3.2 Keyboard Design

Yan et al. (2017) proposed the keyboard design to address the overdosing issue of the mTPI design. Unlike the mTPI design, which divides the toxicity probabilities into three intervals (i.e., underdosing, proper dosing, and overdosing intervals), the keyboard design defines a series of equal-width dosing intervals (referred to as "keys")

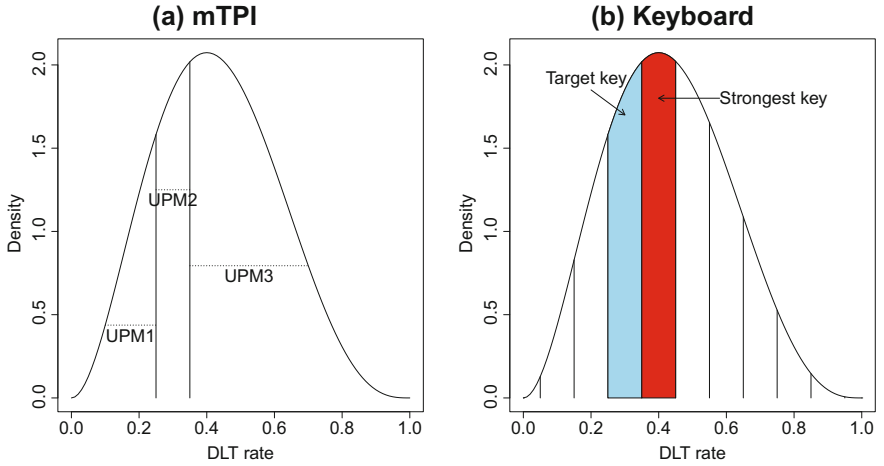


Fig. 8.1 Contrast between the (**a**) mTPI design and (**b**) keyboard design. The curves are the posterior distributions of p_j. To determine the next dose, the mTPI design compares the values of the three UPMs, whereas the keyboard design compares the location of the strongest key with respect to the target key

that correspond to all potential locations of the true toxicity of a particular dose, and uses the key with the highest posterior probability to guide dose escalation and de-escalation. Figure 8.1 contrasts the keyboard design with the mTPI design.

Specifically, the keyboard design starts by specifying a proper dosing interval $\mathscr{I}^* = (\delta_1, \delta_2)$, referred to as the "target key", and then populates this interval toward both sides of the target key, forming a series of keys of equal width that span the range of 0 to 1. For example, given the proper dosing interval or target key of (0.25, 0.35), on its left side, we form 2 keys of width 0.1, i.e., (0.15, 0.25) and (0.05, 0.15); and on its right side, we form 6 keys of width 0.1, i.e., (0.35, 0.45), (0.45, 0.55), (0.55, 0.65), (0.65, 0.75), (0.75, 0.85) and (0.85, 0.95). We denote the resulting intervals/keys as $\mathscr{I}_1, \ldots, \mathscr{I}_K$.

To make the decision of dose escalation and de-escalation, given the observed data $D_j = (n_j, y_j)$ at the current dose level j, the keyboard design identifies the interval $\mathscr{I}_{\max}$ that has the largest posterior probability, i.e.,

$$\mathscr{I}_{\max} = \underset{\mathscr{I}_1,\ldots,\mathscr{I}_K}{\operatorname{argmax}} \{\operatorname{pr}(p_j \in \mathscr{I}_k \mid D_j);\ k = 1, \ldots, K\},$$

which can easily be evaluated based on the posterior distribution of p_j given by Eq. (8.3), assuming that p_j follows a beta-binomial model (8.2). $\mathscr{I}_{\max}$ represents the interval in which the true value of p_j is most likely located, referred to as the "strongest" key by Yan et al. (2017). Graphically, the strongest key is the one with the largest area under the posterior distribution curve of p_j (see Fig. 8.1b). If the strongest key is on the left (or right) side of the target key, that means that the

observed data suggest that the current dose is most likely to represent underdosing (or overdosing), and thus dose escalation (or de-escalation) is needed. If the strongest key is the target key, the observed data support that the current dose is most likely to be in the proper dosing interval, and thus it is desirable to retain the current dose for treating next patient. In contrast, the UPM used by the mTPI design does not have such an intuitive interpretation and tends to distort the evidence for overdosing, as described previously.

Suppose j is the current dose level. The keyboard design determines the next dose as follows.

- If the strongest key is on the left side of the target key, escalate the dose to level $j + 1$.
- If the strongest key is the target key, retain the current dose level j.
- If the strongest key is on the right side of the target key, de-escalate the dose to level $j - 1$.

The trial continues until the prespecified sample size is exhausted, and the MTD is selected based on the isotonic estimates of p_j as described previously.

8.2.3.3 Bayesian Optimal Interval (BOIN) Design

Unlike the mTPI and keyboard designs, which require specifying the prior distribution of p_j and calculating its posterior, the BOIN design is more straightforward. Under the BOIN design, the decision of dose escalation and de-escalation involves only a simple comparison of the observed DLT rate at the current dose with a pair of fixed, prespecified dose escalation and de-escalation boundaries.

Specifically, let $\hat{p}_j = y_j/n_j$ denote the observed DLT rate at the current dose and λ_e and λ_d denote the prespecified dose escalation and de-escalation boundaries. The BOIN design is illustrated in Fig. 8.2 and described as follows:

1. Patients in the first cohort are treated at the lowest dose d_1, or the physician-specified dose.
2. Assuming that the current dose level for treating the latest cohort of patients is j, to assign a dose to the next cohort of patients,
   - if $\hat{p}_j \leq \lambda_e$, escalate the dose to level $j + 1$;
   - if $\hat{p}_j \geq \lambda_d$, de-escalate the dose to level $j - 1$;
   - otherwise, retain the current dose.
3. Repeat step 2 until the maximum sample size N is reached. At that point, select the MTD based on the isotonic estimates of the DLT probabilities as described previously for the mTPI design.

The model-assisted component of the BOIN design is reflected by how the escalation and de-escalation boundaries (λ_e, λ_d) are derived. Specifically, the BOIN assumes that at the current dose level j, the number of patients who experienced DLT (y_j) follows a binomial model:

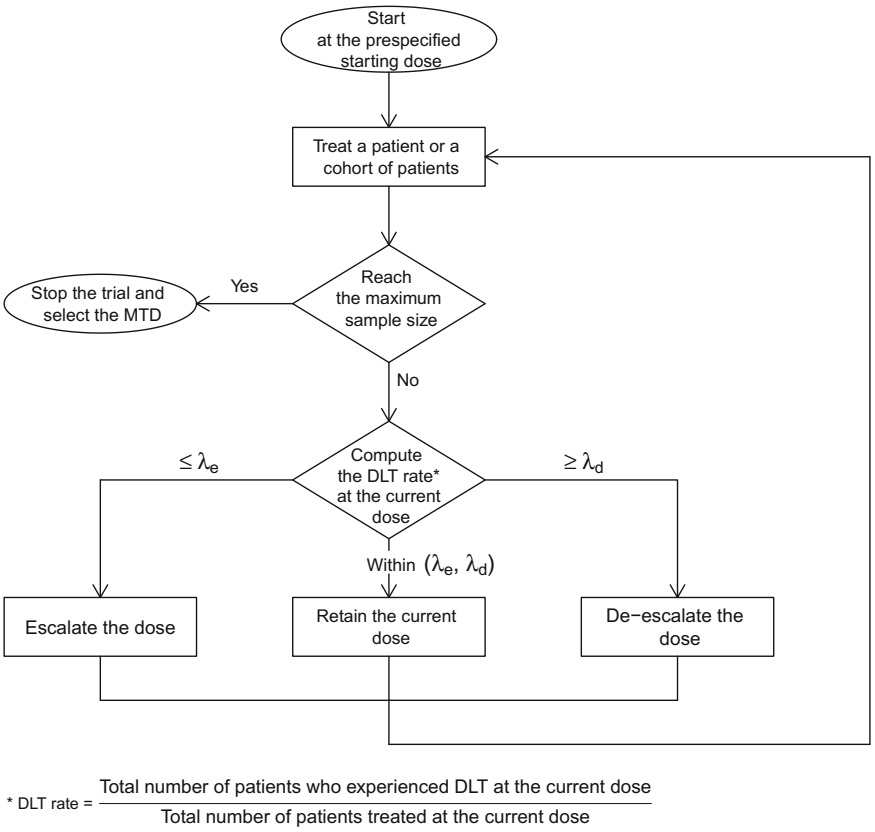


Fig. 8.2 Flowchart of the BOIN design

$$y_j \mid n_j, p_j \sim Binom(n_j, p_j).$$

Under this model assumption, the BOIN minimizes the incorrect decision of dose escalation and de-escalation based on three point hypotheses: $H_1 : p_j = \phi$; $H_2 : p_j = \phi_1$; $H_3 : p_j = \phi_2$, where ϕ_1 denotes the highest DLT rate that is deemed subtherapeutic (i.e., underdosing), such that dose escalation should be made; and ϕ_2 denotes the lowest DLT rate that is deemed overly toxic (i.e., overdosing), such that dose de-escalation is required. The values of ϕ_1 and ϕ_2 can be elicited from physicians. Liu and Yuan (2015) recommended that the default values be used as $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$, which generally yield a design with good operating characteristics. Alternatively, the values of ϕ_1 and ϕ_2 can be calibrated to achieve a particular requirement of the trial at hand. For example, if more conservative dose escalation is required, then setting $\phi_2 = 1.2\phi$ may be adequate. In general, the optimal escalation and de-escalation boundaries (λ_e, λ_d) that minimize the probability of making an incorrect decision of dose assignment arise as

Table 8.3 The escalation/de-escalation boundaries (λ_e, λ_d) under the BOIN design for different target toxicity rates

Boundaries	Target toxicity rate ϕ					
	0.15	0.2	0.25	0.3	0.35	0.4
λ_e	0.118	0.157	0.197	0.236	0.276	0.316
λ_d	0.179	0.238	0.298	0.358	0.419	0.479

Using the default underdosing toxicity rate $\phi_1 = 0.6\phi$ and overdosing toxicity rate $\phi_2 = 1.4\phi$

$$\lambda_e = \frac{\log\left(\frac{1-\phi_1}{1-\phi}\right)}{\log\left\{\frac{\phi(1-\phi_1)}{\phi_1(1-\phi)}\right\}}, \qquad \lambda_d = \frac{\log\left(\frac{1-\phi}{1-\phi_2}\right)}{\log\left\{\frac{\phi_2(1-\phi)}{\phi(1-\phi_2)}\right\}},$$

which minimize the decision error of dose escalation and de-escalation. Liu and Yuan (2015) provided the derivation of these boundaries and showed that the resulting BOIN design has desirable finite-sample and large-sample properties, i.e., long-term memory coherence and consistency.

Table 8.3 provides the values of λ_e and λ_d for commonly used target toxicity rates, obtained with the default values $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$. For example, given the target DLT rate $\phi = 0.3$, the corresponding escalation boundary $\lambda_e = 0.236$ and the de-escalation boundary $\lambda_d = 0.358$. Interestingly, in this case, the 3+3 rule is nested within the BOIN design. That is, escalate/de-escalate/retain the current dose if 0/3 or 2/3 or 1/3 patients have DLT. This feature of the BOIN design links it to established phase I approaches and facilitates communication with clinicians.

It is worth noting that ϕ_1 and ϕ_2 have different interpretations than the proper dosing interval (δ_1, δ_2) used in the mTPI design. As described previously, ϕ_1 and ϕ_2 represent the DLT rates that should be regarded as underdosing and overdosing, respectively; whereas δ_1 and δ_2 represent the range of DLT probabilities that are acceptable. For example, given that the target DLT probability $\phi = 0.25$, setting $\phi_1 = 0.15$ and $\phi_2 = 0.35$ mean that the doses with the DLT rates of 0.15 and 0.35 are respectively regarded as underdosing and overdosing, whereas setting $\delta_1 = 0.15$ and $\delta_2 = 0.35$ means that the dose with a DLT rate between 0.15 and 0.35 is regarded as acceptable. Thus, in general, the value of ϕ_1 should be lower than δ_1 and the value of ϕ_2 should be higher than δ_2. It may be viewed as an advantage that the BOIN design requires users only to specify a DLT rate deemed underdosing (i.e., ϕ_1) and a DLT rate deemed overdosing (i.e., ϕ_2), rather than binning the DLT rate into the proper dosing interval and improper dosing intervals as required by the mTPI and keyboard designs. Binning the DLT rate into intervals causes the "discontinuity" dilemma. For example, by setting $(\delta_1, \delta_2) = (0.15, 0.30)$, the mTPI defines $p_j = 0.299$ as proper dosing and $p_j = 0.301$ as overdosing, which may look odd in practice as these two DLT rates are virtually the same.

The BOIN design and the CRM design are similarly flexible. The BOIN design can target any prespecified DLT rate. For instance, for some cancer populations for

whom there is no effective treatment, a target DLT rate higher than 0.3 may be an acceptable trade-off to achieve higher treatment efficacy; while for other cancer populations, a lower target DLT rate, e.g., 0.15 or 0.2, may be more appropriate. In addition, unlike the 3+3 design, for which the dose escalation and de-escalation decisions can be made only when we have 3 or 6 evaluable patients, the BOIN design does not require a fixed cohort size and allows for decision making at any time during the trial by comparing the observed DLT rate at the current dose with the escalation and de-escalation boundaries. Decisions regarding dose escalation and de-escalation can be made at any time as long as we can calculate the DLT rate at the current dose. Such flexibility has important practical utility and implications. It allows clinicians to "adaptively" change the cohort size during the course of the trial to achieve certain design goals. For example, to shorten the trial duration and reduce the sample size, clinicians often prefer to use a cohort size of 1 for the initial dose escalation and then switch to a cohort size of 3 after observing the first DLT, as with the commonly used accelerated titration design (ATD) (Simon et al. 1997). Such an accelerated titration can be easily and seamlessly performed using the BOIN design by simply switching the cohort size from 1 to 3 when the first DLT is observed. Unlike the ATD, which combines two independent empirical rules (the accelerated titration rule and the 3+3 rule), in an ad hoc way, the BOIN design achieves the same design goal under a single, coherent framework with assured statistical properties.

Operatively and conceptually, the BOIN design is simpler and more transparent than the mTPI/mTPI-2 and keyboard designs. To make the decision of dose escalation and de-escalation, the BOIN design simply compares the observed DLT rate $\hat{p}_j$ with a pair of fixed, prespecified escalation and de-escalation boundaries (λ_e, λ_d); whereas the keyboard and mTPI/mTPI2 designs require calculating the posterior distribution and identifying the "strongest" key and UPM, respectively, for each possible outcome data (y_j, n_j), though these evaluations and corresponding decision boundaries can be calculated prior to the onset of the trial. In addition, thanks to the feature that the BOIN design guarantees de-escalating the dose when the observed toxicity rate $\hat{p}_j$ is higher than the de-escalation boundary λ_d, it is particularly easy for clinicians and regulatory agents to assess the safety of a trial using the BOIN design. For example, given a target DLT rate $\phi = 0.25$, we know a priori that a phase I trial using the BOIN design guarantees de-escalating the dose if the observed DLT rate is higher than 0.298 (with the default values of ϕ_1 and ϕ_2). Accordingly, the BOIN design also allows users to easily calibrate the design to satisfy a specific safety requirement mandated by regulatory agents through choosing an appropriate target DLT rate ϕ or ϕ_2. For example, supposing for a phase I trial with a new compound, the regulatory agent mandates that if the observed toxicity rate is higher than 0.25, the dose must be de-escalated. We can easily fulfill that requirement by setting the target DLT rate $\phi = 0.20$, under which the BOIN automatically guarantees de-escalating the dose if the observed toxicity rate $\hat{p}_j > \lambda_d = 0.238$. If needed, the de-escalation boundary λ_d can be further fine tuned by calibrating the value of ϕ_2. Such flexibility and transparency renders the BOIN design an important advantage over the mTPI/mTPI-2 and keyboard designs in practice.

8.2.4 Comparison of Phase I Designs

8.2.4.1 Simulation Settings

We considered target DLT rates of $\phi = 0.2$ and $\phi = 0.3$ with $J = 6$ dose levels. Under each setting, 1000 dose-toxicity scenarios were randomly generated using the method of Clertant and O'Quigley (2017). Under each scenario, we simulated 2000 trials. Figure 8.3 displays 50 randomly selected scenarios with $\phi = 0.20$ and $J = 6$. These exhibit a variety of dose-toxicity curve shapes and spacings. We compared the CRM, mTPI, keyboard and BOIN designs. For the CRM, we used (0.032, 0.095, 0.200, 0.332, 0.470, 0.596) as the skeleton. For the mTPI and keyboard designs, we used the recommended default values $\delta_1 = \phi - 0.05$ and $\delta_2 = \phi + 0.05$. For the BOIN design, we set $\phi_1 = 0.6\phi$ and $\phi_2 = 1.4\phi$, which are the recommended default values. We set the cohort size equal to 1 and the maximum sample size equal to 36. A more comprehensive and complete comparison of these designs is provided by Zhou et al. (2018a, b).

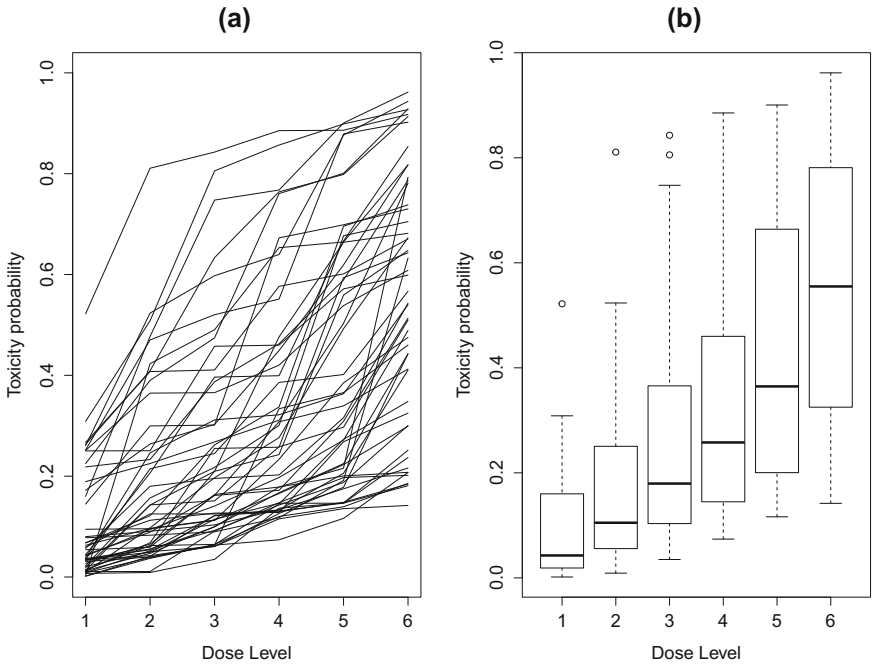


Fig. 8.3 Panel **a** shows 50 randomly selected dose-toxicity curves, and panel **b** shows the distribution of the toxicity probabilities by dose level from the 1000 scenarios with 6 dose levels

8.2.4.2 Performance Metrics

For each of the 1000 scenarios, we calculated the following metrics:

MTD selection

- The percentage of correct selection (PCS), which we define as the percentage of simulated trials in which the correct dose is selected as the MTD.
- The PCS within a 5% acceptable region, which we define as the percentage of simulated trials in which the dose selected as the MTD has a DLT rate that lies in the interval $[\phi - 0.05, \phi + 0.05]$.

Patient allocation

- The average percentage of patients in the simulated trials who are assigned to the MTD.
- The average percentage of patients in the simulated trials who are assigned to a dose with a DLT rate that lies in the interval $[\phi - 0.05, \phi + 0.05]$.

Overdose control

- The average number of patients in the simulated trials who are assigned to a dose that is above the MTD.
- The risk of overdosing, which we define as the percentage of simulated trials in which a large percentage of patients (e.g., 80%) are assigned to a dose that is above the MTD. This metric quantifies how likely a particular design is to overdose a large percentage of patients.

8.2.4.3 Results

Table 8.4 summarizes the average performance of the CRM, mTPI, BOIN and keyboard designs. In general, the CRM, BOIN and keyboard designs provide comparable, excellent operating characteristics, and each outperforms the mTPI design with higher probability of correctly selecting the MTD and less likelihood of overdosing patients.

MTD Selection

Figure 8.4 shows the results for PCS and PCS within a 5% acceptable region for the mTPI, BOIN and keyboard designs, with respect to the CRM. Each boxplot reflects the distribution of the corresponding metric across the 1000 scenarios, and the red $\times$ reflects the average. As an example, the top-left panel of Fig. 8.4 shows a boxplot of the PCS difference between mTPI and CRM, between BOIN and CRM, and between keyboard and CRM when $\phi = 0.20$. For mTPI versus CRM, most of the data points are negative, which indicates that the CRM tends to outperform mTPI. For BOIN and keyboard versus CRM, respectively, most of the data points are close to zero, which indicates that the BOIN and keyboard designs tend to perform similarly to the CRM. For PCS within 5% (bottom panels), we see a similar pattern. As evidenced by the

Table 8.4 Average performance of the CRM, BOIN, mTPI and keyboard designs across 1000 scenarios with 6 dose levels

Performance metric	Target $\phi = 0.20$				Target $\phi = 0.30$			
	CRM	BOIN	mTPI	Keyboard	CRM	BOIN	mTPI	Keyboard
PCS (%)	50.2	51.9	44.0	51.4	51.0	51.8	49.0	51.8
PCS within 5% (%)	61.3	61.8	52.1	61.3	59.7	59.9	56.8	59.9
Patients treated at MTD (%)	39.1	39.3	36.7	39.1	39.9	39.4	39.4	39.3
Patients treated within 5% (%)	48.6	48.0	44.2	47.9	47.3	46.4	46.3	46.3
Number of patients treated above MTD	6.2	7.6	7.4	7.3	7.5	7.9	8.9	7.8
Risk of overdosing 80% (%)	6.6	7.4	15.6	7.2	9.0	8.4	16.0	8.0

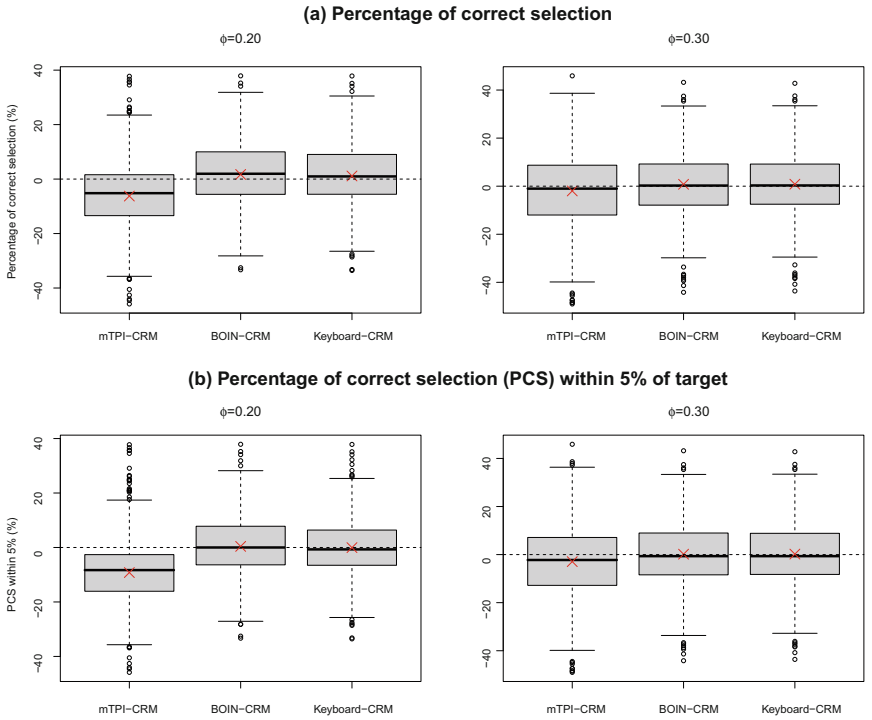


Fig. 8.4 Boxplot of the difference in the percentage of correct selection (PCS) of the MTD and the PCS of the doses within 5% of the target for mTPI versus CRM, BOIN versus CRM and keyboard versus CRM under 1000 scenarios with 6 dose levels. The red $\times$ reflects the average difference

right panels of Fig. 8.4, when $\phi = 0.30$ the CRM, BOIN and keyboard designs have comparable PCS and PCS within 5%, and all outperform mTPI, though to a lesser extent than when $\phi = 0.20$.

Patient Allocation

Figure 8.5 shows the results for the average percentage of patients who are assigned to the MTD, and the average percentage of patients treated at the doses within the 5% acceptable region of the target DLT rate, respectively, for $\phi = 0.20$ and $\phi = 0.30$ with $J = 6$ doses. When $\phi = 0.20$, CRM, BOIN and keyboard are comparable, and all tend to outperform mTPI. When $\phi = 0.30$, all four designs are comparable with respect to the two metrics.

Overdose Control

Overdose control is important for protecting patients from overly toxic doses. The upper panel of Fig. 8.6 shows our results for the number of patients treated above the

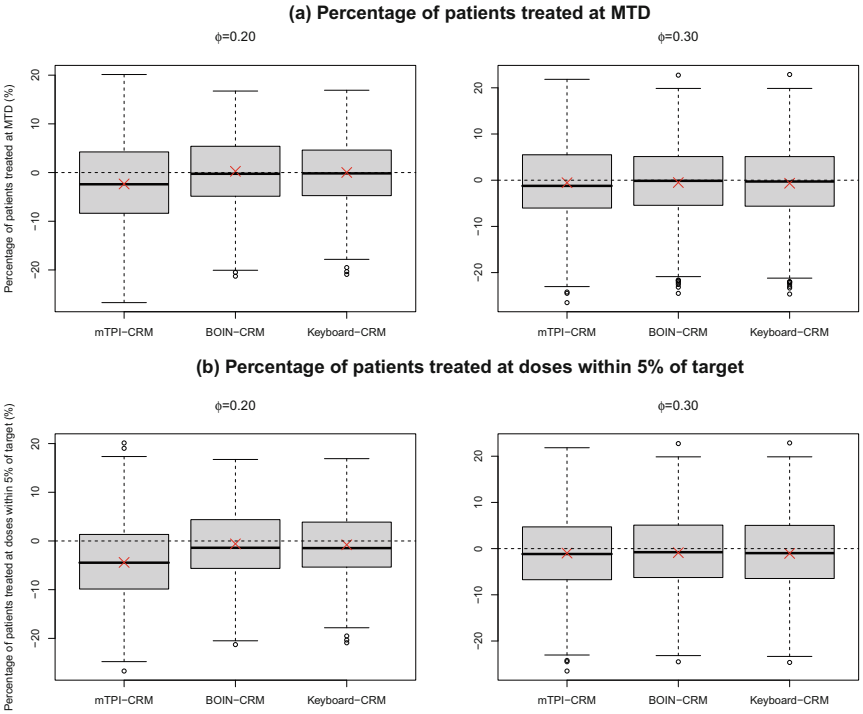


Fig. 8.5 Boxplot of the difference in the percentage of patients treated at the MTD and the percentage of patients treated at doses within 5% of the target for mTPI versus CRM, BOIN versus CRM and keyboard versus CRM under 1000 scenarios with 6 dose levels. The red × reflects the average difference

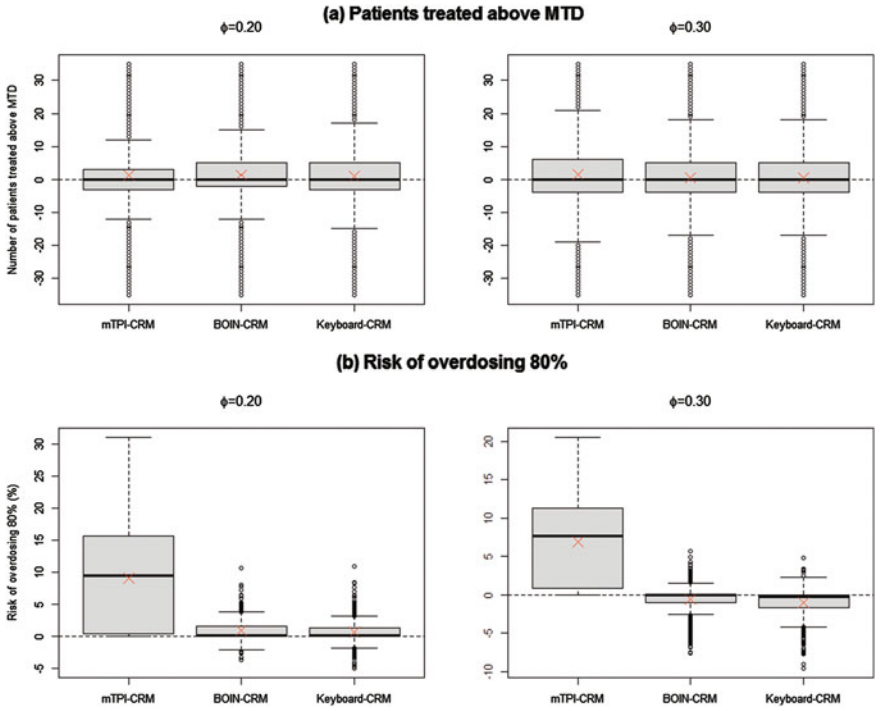


Fig. 8.6 Boxplot of the difference in the number of patients treated above the MTD and the risk of overdosing at least 80% of patients for mTPI versus CRM, BOIN versus CRM and keyboard versus CRM under 1000 scenarios with 6 dose levels. The red $\times$ reflects the average difference

MTD in each simulated trial under 1000 scenarios. Generally speaking, the four designs have comparable performance. Table 8.4 shows that, compared with the BOIN, mTPI and keyboard designs, the CRM tends to assign fewer patients to doses that are above the MTD when $\phi = 0.20$. When $\phi = 0.30$, the CRM, BOIN and keyboard perform similarly and tend to assign fewer patients to doses that are above the MTD than the mTPI. The lower panel of Fig. 8.6 shows the results for the risk of overdosing at least 80% of the patients. The difference between the designs for this safety metric is more striking than for the other metrics. For instance, the difference between mTPI and CRM in the risk of overdosing at least 80% of the patients is greater than zero in every scenario, which indicates that the CRM is always safer than the mTPI design. In contrast, the CRM, BOIN and keyboard designs are comparable regarding the risk of overdosing at least 80% of the patients. When $\phi = 0.20$, the CRM is slightly safer than BOIN and keyboard, whereas when $\phi = 0.30$, BOIN and keyboard are slightly safer than the CRM. Table 8.4 shows that the average risk of overdosing at least 80% of the patients is substantially lower for the CRM, BOIN and keyboard designs than for the mTPI design.

Table 8.5 Escalation and de-escalation rules for the mTPI, BOIN and keyboard designs under their default settings for a target toxicity rate of $\phi = 0.2$

	Number of patients treated at the current dose															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
mTPI design																
Escalate if number of DLTs $\leq$	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
De-escalate if number of DLTs $\geq$	1	2	2	2	3	3	4	4	4	5	5	5	5	6	6	6
BOIN design																
Escalate if number of DLTs $\leq$	0	0	0	0	0	0	1	1	1	1	1	1	2	2	2	2
De-escalate if number of DLTs $\geq$	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
Keyboard design																
Escalate if number of DLTs $\leq$	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2	2
De-escalate if number of DLTs $\geq$	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4

The reason the mTPI design is more likely than the other designs to overdose at least 80% of the patients can be seen through the dose escalation and de-escalation rules for the three model-assisted designs reported in Table 8.5. When the target is $\phi = 0.2$, the default BOIN, mTPI and keyboard designs use different thresholds for dose escalation and de-escalation. In particular, compared to the BOIN and keyboard designs, the mTPI design uses a less aggressive escalation rule and a less conservative de-escalation rule. For example, suppose 16 patients have been assigned to the current dose, the mTPI design escalates the dose if 1 or fewer DLTs have been observed and only de-escalates the dose if 6 or more DLTs have been observed. This is in contrast to the BOIN and keyboard designs, which escalate the dose if 2 or fewer DLTs have been observed and de-escalate the dose if 4 or more DLTs have been observed. Consequently, the mTPI design tends to get stuck at a particular dose once 10 or more patients have been treated at that dose. In particular, if the dose at which the mTPI design happens to become stuck is above the MTD, a large percentage of patients will be overdosed.

8.2.5 Software

Shiny apps to implement the aforementioned designs, including CRM, BMA-CRM, BOIN, and keyboard designs, are freely available at http://www.trialdesign.org. Figure 8.7 shows the interface of the Shiny app for the BOIN design, which allows

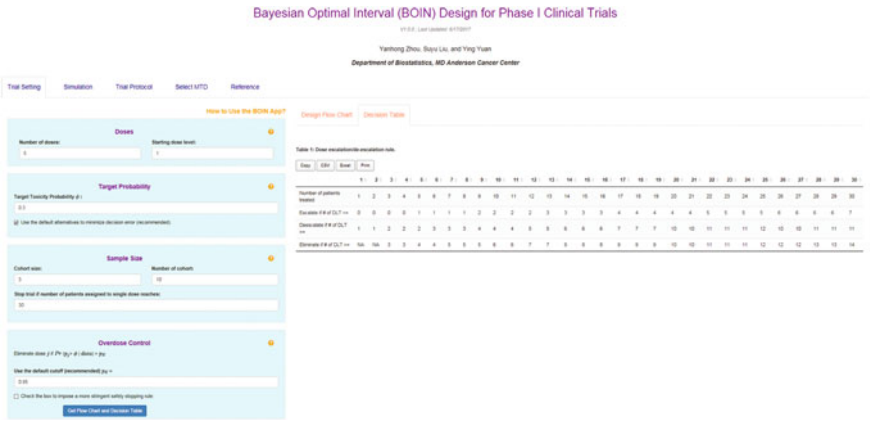


Fig. 8.7 Shiny app for implementing the BOIN design for single agent trial

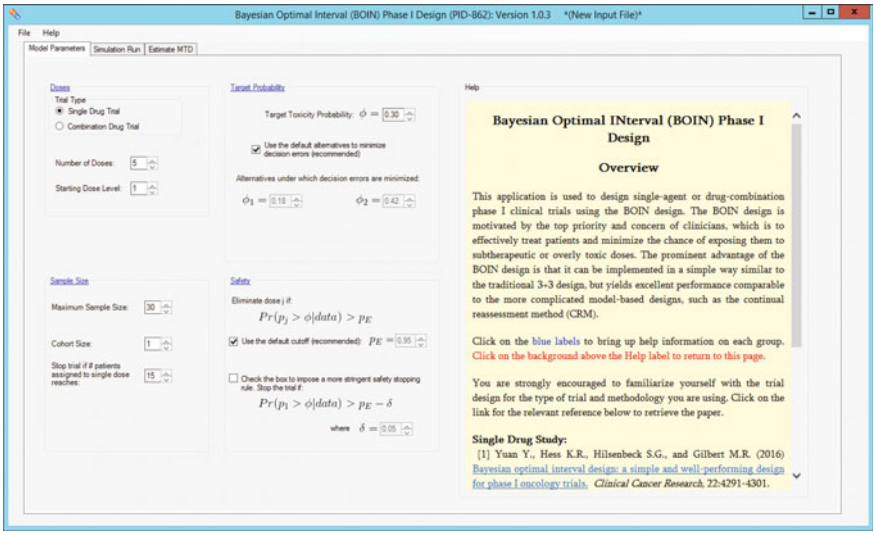


Fig. 8.8 Windows® desktop program for the BOIN design

users to generate the decision table for dose escalation and de-escalation, perform a simulation study to generate the operating characteristics of the design, and create the template for trial protocol preparation. To facilitate the use of the novel designs, the BOIN design is also available in other easy-to-use forms, including a stand-alone Windows® desktop program freely available from https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=99, and the R package "BOIN" available from CRAN. The software comes with detailed documents and provides step-by-step instructions on how to use it to design phase I trials. Figure 8.8 show the Windows® desktop program for the BOIN design.

8.3 Drug Combination Trial

Drug combination therapy provides an important approach for treating difficult diseases such as cancer. The objectives of using a combination of drugs are to induce a synergistic treatment effect, increase the joint dose intensity with non-overlapping toxicities, and target various tumor cell susceptibilities and disease pathways.

A major challenge in designing drug combination trials is that such combinations are only *partially ordered* according to their toxicity probabilities. Consider a trial combining J doses of agent A, denoted as $A_1 < A_2 < \ldots < A_J$, and K doses of agent B, denoted as $B_1 < B_2 < \ldots < B_K$. Let A_jB_k denote the combination of A_j and B_k, and p_{jk} denote the DLT rate for A_jB_k. It is typically reasonable to assume that when the dose of one agent (say agent A) is fixed, the toxicity of the combination increases as the dose of the other agent increases (i.e., agent B). In other words, as shown in Fig. 8.9, in the dose matrix, the rows and columns are ordered, with the DLT rate increasing along with the dose. However, in other directions of the dose matrix (e.g., along the diagonals from the upper left corner to the lower right corner), the toxicity order is unknown due to unknown drug-drug interactions. For example, between A_2B_2 and A_1B_3, we do not know which drug is more toxic because the first combination has a higher dose of agent A whereas the second combination has a higher dose of agent B. Thus, we cannot fully rank $J \times K$ combinations from low to high in terms of their DLT rates. This is distinctly different from single-agent trials, for which the dose can be unambiguously ranked assuming that higher dosage yields higher DLT rate. The implication of such a partial ranking is that conventional single-agent dose-finding designs cannot be directly used for finding the MTD in drug combination trials.

Another challenge for combination trials is the existence of the *MTD contour* in the two-dimensional dose space, as shown in Fig. 8.10. As a result, multiple MTDs may exist in the $J \times K$ dose matrix. The implication of the MTD contour is that when

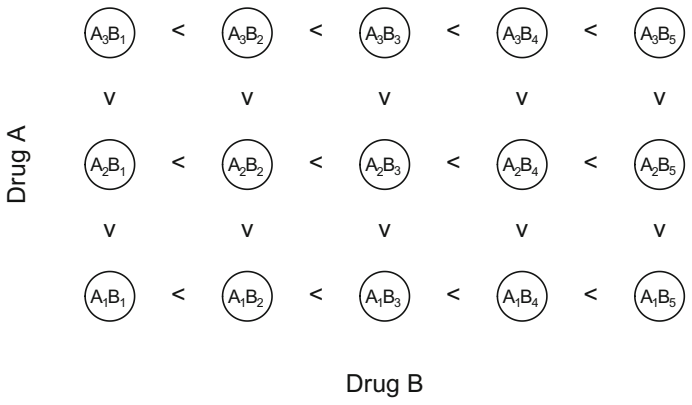


Fig. 8.9 Partial order in toxicity for drug combinations

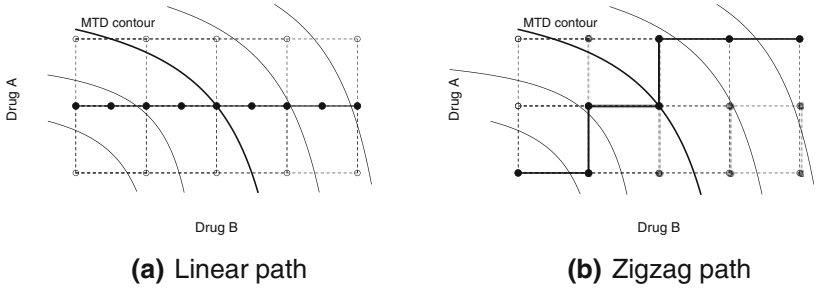


Fig. 8.10 Illustration of the linearization approach to find a single MTD for drug combination trials. The lattices of dotted lines denote the dose combination matrix; solid curved lines indicate the toxicity contours; and the solid line with solid circles indicates the linear path for dose finding

designing a drug combination trial, the first and most important question requiring careful consideration is

Are we interested in finding one MTD or multiple MTDs?

As we describe below, the answer to this question determines the choice of different design strategies for drug combination trials. This important issue, unfortunately, is largely overlooked by existing trial designs.

8.3.1 Combination Trials to Find One MTD

8.3.1.1 Model-Based Designs

Numerous designs have been proposed to find a single MTD for drug combinations. For example, Conaway et al. (2004) proposed a drug combination dose-finding method based on the order of the restricted inference. Yin and Yuan (2009a, 2009b) proposed Bayesian dose-finding designs based on latent contingency tables (Yin and Yuan 2009a) and a copula-type model (Yin and Yuan 2009b) for drug combination trials. Braun and Wang (2010) developed a dose-finding method based on a Bayesian hierarchical model. Wages et al. (2011) extended the CRM (Pepe et al. 1990) based on partial ordering of the dose combinations. Braun and Jia (2013) generalized the CRM to handle drug combination trials. Riviere et al. (2014) proposed a Bayesian dose-finding design based on the logistic model. Cai et al. (2014) and Riviere et al. (2015) proposed Bayesian adaptive designs for drug combination trials involving molecularly targeted agents. Albeit very different, most of these designs adopt a common dose-finding strategy similar to the CRM: devise a model to describe the dose-toxicity surface and then, based on the accumulating data, continuously update the model estimate and make the decision of dose assignment for the new patient,

typically by assigning the new patient to the dose for which the estimated DLT rate is closest to the target DLT rate.

Although these designs perform reasonably well, they are rarely used in practice for several reasons. First, these designs are statistically and computationally complicated, leading many practitioners to perceive that the decisions of dose allocation arise from a "black box". Lack of easy-to-use software further hinders the adoption of these designs in practice. Robustness is another potential issue for model-based drug combination trial designs. As the models used in the drug-combination designs are more complicated than the CRM, the designs are more vulnerable to model misspecification, and also the dose-finding scheme is much more likely to become stuck at local "suboptimal" doses. Some strategies [e.g., giving high priority to exploring new doses (Cai et al. 2014), and randomization (Riviere et al. 2015)] have been proposed to alleviate this issue, but given the small sample size of early phase trials, this remains an issue that affects the robustness of drug combination trial designs. The robustness of the model-based drug combination trial designs warrants further research. Because of the aforementioned issues, we do not discuss these model-based approaches further. Instead, in what follows, we focus on two simple and robust approaches that can be easily implemented using existing Shiny apps or a Windows® desktop program, making them more likely to be used in practice.

8.3.1.2 Linearization Approach

When the goal is to find a single MTD, a much simpler, robust approach to drug combination trials is available. The key observation is that there is no need to search the whole (partially ordered) dose matrix. As demonstrated by Fig. 8.10, we can select a certain ordered path (i.e., a sequence of combinations), which starts from a low dose combination (e.g., lower left corner) and ends at a high dose combination (e.g., upper right corner), to find the MTD. This approach, which we call "*linearization*", has been widely used in practice to design drug combination trials. One may argue that compared to searching the dose matrix, the linearization approach is more likely to miss the MTD because the MTD is less likely to be in the selected linear path than in the whole dose matrix. That occurs simply because the dose matrix contains more doses to be investigated. For example, a 4×4 drug combination matrix contains 16 investigational doses. In the linearization approach, if we specify the same number of, say 16, doses (on a finer grid), there is little reason for the linear path to be less likely than the dose matrix to contain the MTD. This is because in principle, as the number of doses increases, the linear path will eventually hit the MTD contour (see Fig. 8.10). Actually, given a prespecified $J \times K$ dose matrix, there is also no guarantee that it contains the MTD. The same argument applies to traditional single-agent dose finding as well (In this chapter, we use the terms one-dimensional dose finding and single-agent dose finding to indicate the same thing). Chu et al. (2016) proposed a method to adaptively add new doses when the trial data indicate that none of the prespecified doses are close to the target DLT rate. That method can be used

with the linearization approach to address the concern that none of the doses is close to the MTD.

The beauty of linearization is that it converts a complex, partially ordered dose combination matrix into a sequence of ordered combinations. Therefore, the existing single-agent dose-finding methods, for example, the BOIN, keyboard or CRM, can be directly used to find the MTD. Depending on the clinical setting, different linearization paths can be used. For example, if drug A is the standard treatment and serves as the backbone of the combination treatment, we may prefer to fix A at its standard dose and vary the dose of drug B (see Fig. 8.10a). In other settings, such as when two drugs are similarly important, we may prefer to alternatively increase the doses of A and B, which results in a zigzag line in the dose surface (see Fig. 8.10b).

8.3.1.3 BOIN Drug-Combination Design

Unlike the linearization approach that requires users to select a specific ordered path in the dose combination matrix, the BOIN drug-combination design (Lin and Yin 2015) provides a simple, well-performing method to find a single MTD directly in the two-dimensional dose matrix. The BOIN drug-combination design makes the decision of dose escalation/de-escalation based on the same rule as the single-agent BOIN design described previously. The only difference is that, in combination trials, when we decide to escalate or de-escalate the dose, there is more than one neighboring dose to which we can move. For example, when we escalate/de-escalate the dose, we can escalate/de-escalate either the dose of drug A or the dose of drug B. The BOIN drug-combination design makes this choice based on $\text{pr}(p_{jk} \in (\lambda_e, \lambda_d)|\text{data})$, which measures the likelihood of a dose combination being located within (λ_e, λ_d) given the observed data, where λ_e and λ_d are the escalation and de-escalation boundaries same as those for single-agent BOIN design, described previously. The beta-binomial model described above can be easily used to evaluate $\text{pr}(p_{jk} \in (\lambda_e, \lambda_d)|\text{data})$.

Let $\hat{p}_{jk} = y_{jk}/n_{jk}$ denote the observed DLT rate at dose combination A_jB_k, where y_{jk} and n_{jk} denote the number of toxicities and patients treated at A_jB_k, respectively. Define an admissible dose escalation set as $\mathscr{A}_E = \{A_{j+1}B_k, A_jB_{k+1}\}$ and an admissible dose de-escalation set as $\mathscr{A}_D = \{A_{j-1}B_k, A_jB_{k-1}\}$. The BOIN drug-combination design can be described as follows.

1. Patients in the first cohort are treated at the lowest dose combination A_1B_1 or a prespecified dose combination.
2. Suppose the current cohort is treated at dose combination A_jB_k, then to assign a dose to the next cohort of patients, we follow these rules.
   - If $\hat{p}_{jk} \le \lambda_e$, escalate the dose to the combination that belongs to $\mathscr{A}_E$ and has the largest value of $\text{pr}\{p_{j'k'} \in (\lambda_e, \lambda_d)|\text{data}\}$.
   - If $\hat{p}_{jk} \ge \lambda_d$, de-escalate the dose to the combination that belongs to $\mathscr{A}_D$ and has the largest value of $\text{pr}\{p_{j'k'} \in (\lambda_e, \lambda_d)|\text{data}\}$.
   - Otherwise, if $\lambda_e < \hat{p}_{jk} < \lambda_d$, stay at the same combination A_jB_k.

3. This process is continued until the maximum sample size is reached or the trial is terminated because of excessive toxicity.

During dose escalation and de-escalation, if the two combinations in $\mathscr{A}_E$ or $\mathscr{A}_D$ have the same value of $\text{pr}\{p_{j'k'} \in (\lambda_e, \lambda_d)|\text{data}\}$, we randomly choose one with equal probability. If no dose combination exists in the sets of $\mathscr{A}_E$ and $\mathscr{A}_D$ (i.e., we are at the boundaries of the dose matrix), we retain the current dose combination. After the trial is completed, the MTD is selected as the dose combination with the estimated DLT rate closest to ϕ. The estimates of DLT rates are obtained using isotonic regression as described previously, but in a matrix form. More details on the BOIN drug-combination design can be found in Lin and Yin (2015).

8.3.2 Combination Trials to Find Multiple MTDs

The primary motivation for combining drugs is to achieve synergistic treatment effects. Because of the existence of the MTD contour and the fact that doses on the MTD contour may have different efficacy due to drug-drug interactions, for many drug combination trials, it is of intrinsic interest to find multiple MTDs. The efficacy of the MTDs can be evaluated in subsequent phase II trials or simultaneously in phase I-II trials. Given a prespecified $J \times K$ dose matrix, finding the MTD contour is equivalent to finding an MTD, if it exists, in each row of the dose matrix. Without loss of generality, we assume that $J \leq K$. That is, drug B has more dose levels than drug A.

Finding the MTD contour is substantially more challenging than finding a single MTD. This is because in order to find all MTDs in the dose matrix, we must explore the whole dose matrix using the limited sample size that is a characteristic of phase I trials; otherwise, we risk missing some MTDs. In contrast to numerous drug combination designs that have been proposed for finding a single MTD, a very limited number of designs for finding the MTD contour have been proposed. Thall et al. (2003) proposed a drug combination design to find three MTDs, but that design assumes that the doses are continuous and can be freely changed during the trial, which is not common in practice. Wang and Ivanova (2005) proposed a design to find the MTD contour based on a parametric model, assuming that the logarithm of the DLT rate of a drug combination is a linear function of the doses of the two drugs. Yuan and Yin (2008) proposed a sequential dose-finding method that converts the task of finding the MTD contour into a series of easier one-dimensional dose-finding problems. Mander and Sweeting (2015) proposed a product of independent beta probabilities escalation (PIPE) design to find the MTD contour based on Bayesian model averaging, without assuming a parametric form on the dose-toxicity curve. Zhang and Yuan (2016) extended the approach of Yuan and Yin (2008) and proposed a so-called waterfall design to incorporate some practical considerations. Because the waterfall design is easy to implement, has good performance and easy-to-use software, we focus on the waterfall design hereafter.

The basic idea of the waterfall design is straightforward: divide the two-dimensional dose-finding problem into a series of simpler one-dimensional dose-finding problems that can be easily solved by existing single-agent dose-finding methods, where each one-dimensional dose-finding process is known as a "subtrial". As illustrated in Fig. 8.11, the waterfall design partitions the $J \times K$ dose matrix into J subtrials, within which the doses are fully ordered. These subtrials are conducted sequentially from the top of the matrix to the bottom, which is why we refer to the design as the waterfall design. The goal of the design is to find the MTD contour, which is equivalent to finding the MTD, if it exists, in each row of the dose matrix. The waterfall design can be described as follows:

1. Divide the $J \times K$ dose matrix into J subtrials $S_J, \ldots, S_1$, according to the dose level of drug A:

$$\begin{aligned} S_J &= \{A_1B_1, \ldots, A_JB_1, A_JB_2, \ldots, A_JB_K\}, \\ S_{J-1} &= \{A_{J-1}B_2, \ldots, A_{J-1}B_K\}, \\ S_{J-2} &= \{A_{J-2}B_2, \ldots, A_{J-2}B_K\}, \\ &\ldots \\ S_1 &= \{A_1B_2, \ldots, A_1B_K\}. \end{aligned}$$

   Note that subtrial S_J also includes lead-in doses $A_1B_1, A_2B_1, \ldots, A_JB_1$ (the first column of the dose matrix) to impose the practical consideration that the trial starts at the lowest dose. Within each subtrial, the doses are fully ordered with monotonically increasing toxicity.
2. Conduct the subtrials sequentially using the BOIN design (or other single-agent dose-finding method) as follows:
   (i) Conduct subtrial S_J, starting from the lowest dose combination A_1B_1, to find the MTD. We call the dose selected by the subtrial a "candidate MTD" to highlight that the dose selected by the individual subtrial may not be the "final" MTD that we select at the end of the trial. The final MTD selection is based on the data collected from all the subtrials. The objective of finding the candidate MTD is to determine which subtrial to conduct next and the corresponding starting dose.
   (ii) Assuming that subtrial S_J selects dose $A_{j^*}B_{k^*}$ as the candidate MTD, next, conduct subtrial S_{j^*-1} with the starting dose $A_{j^*-1}B_{k^*+1}$. That is, the next subtrial to be conducted is the one with the dose of drug A that is one level lower than the candidate MTD found in the previous subtrial. After identifying the candidate MTD of subtrial S_{j^*-1}, the same rule is used to determine the next subtrial and its starting dose. See Fig. 8.11 for an example.
   (iii) Repeat step (ii) until subtrial S_1 is completed.
3. Estimate the DLT rate p_{jk} based on the toxicity data collected from all the subtrials using matrix isotonic regression (Gordon et al. 1984). For each row of the dose matrix, select the MTD as the dose combination that has the estimate

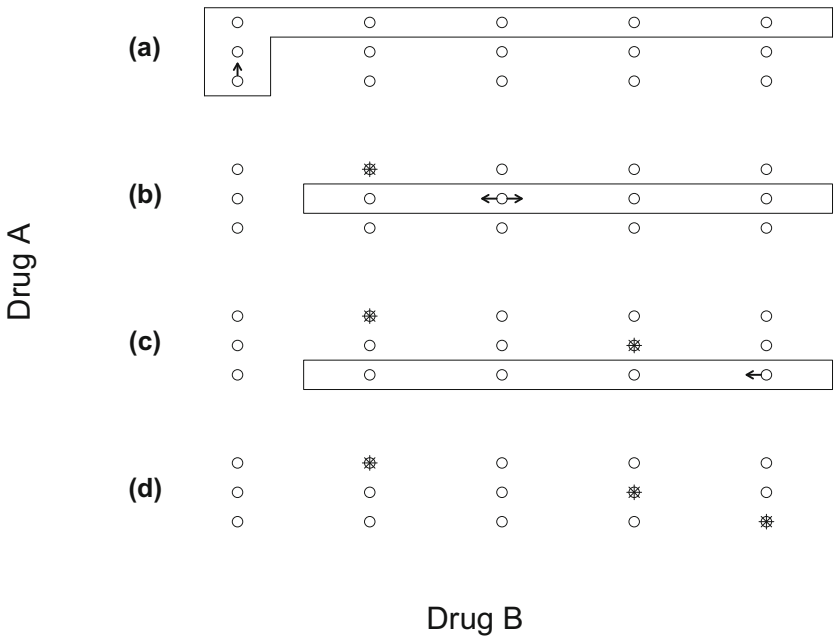


Fig. 8.11 Illustration of the waterfall design for a 3×5 combination trial. The doses in the rectangle form a subtrial, and the asterisk denotes the candidate MTD for the subtrial. As shown in panel (**a**), the trial starts by conducting the first subtrial with the starting dose A_1B_1. After the first subtrial identifies A_3B_2 as the candidate MTD, we then conduct the second subtrial with the starting dose A_2B_3 [see panel (**b**)]. After the second subtrial identifies A_2B_4 as the candidate MTD, we conduct the third subtrial with the starting dose A_1B_5 [see panel (**c**)]. After all subtrials are completed, we select the MTD contour based on the data from all subtrials, as shown in panel (**d**)

of DLT rate that is closest to the target DLT rate ϕ unless all combinations in that row are overly toxic.

The waterfall design conducts the subtrials sequentially such that the results of each subtrial are used to inform the design (e.g., the dose range and the starting dose) of subsequent subtrials. Such information borrowing allows the design to explore the two-dimensional dose space efficiently using a limited sample size, and decreases the chance of overdosing or underdosing patients. For example, in step 2, the reason that subtrial S_{j^*-1} starts with dose $A_{j^*-1}B_{k^*+1}$ rather than the lowest dose in that subtrial (i.e., $A_{j^*-1}B_2$) is that $A_{j^*-1}B_{k^*+1}$ is the lowest dose that is potentially located at the MTD contour. Starting from $A_{j^*-1}B_{k^*+1}$ allows us to quickly reach the MTD. Using Fig. 8.11 as an example, the first subtrial S_3 identified the dose A_3B_2 as the MTD, and thus the second subtrial S_2 starts from the dose A_2B_3. It is not desirable to start from the lowest dose A_2B_2 because the partial ordering informs us that A_2B_2 is below the MTD. Starting at the lowest dose in this example wastes patient resources and exposes patients to low doses that may be subtherapeutic.

8.3.3 Software

The BOIN drug-combination and waterfall designs can be easily implemented using the Shiny app available at http://www.trialdesign.org, or the Windows® desktop program freely available at https://biostatistics.mdanderson.org/softwaredownload/SingleSoftware.aspx?Software_Id=99. These two software applications have intuitive graphical user interfaces and rich documents to help the users. These applications allow users to perform simulations to obtain the operating characteristics of the design, generate the protocol template, and conduct the trial in real time. For users who are comfortable with programming language R, the package "BOIN" is freely available from CRAN to implement the BOIN drug-combination and waterfall designs. The manual for the package can be found at https://cran.r-project.org/web/packages/BOIN/index.html, and a statistical tutorial for using the package to design drug combination trials can be found at http://odin.mdacc.tmc.edu/~yyuan/index_code.html

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1973). Statistical inference under order restrictions: The theory and application of isotonic regression. *International Statistical Review*, *41*(3).

Braun, T. M., & Jia, N. (2013). A generalized continual reassessment method for two-agent phase I trials. *Statistics in Biopharmaceutical Research*, *5*, 105–115.

Braun, T. M., & Wang, S. F. (2010). A hierarchical Bayesian design for phase I trials of novel combinations of cancer yherapeutic agents. *Biometrics*, *66*(3), 805–812.

Cai, C. Y., Yuan, Y., & Ji, Y. (2014). A Bayesian phase I/II design for oncology clinical trials of combining biological agents. *Journal of the Royal Statistical Society: Series C*, *63*, 159–173.

Chu, Y., Pan, H., & Yuan, Y. (2016). Adaptive dose modification for phase I clinical trials. *Statistics in Medicine*, *35*(20), 3497–3508.

Clertant, M., & Quigley, J.O. (2017). Semiparametric dose finding methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. https://doi.org/10.1111/rssb.12229

Conaway, M. R., Dunbar, S., & Peddada, S. D. (2004). Designs for single- or multiple-agent phase I trials. *Biometrics*, *60*(3), 661–669.

Gordon, B., Richard, D., Carolyn, P., & Tim, R. (1984). Isotonic regression in two independent variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *33*(3), 352–357.

Iasonos, A., & O'Quigley, J. (2014). Adaptive dose-finding studies: A review of model-guided phase I clinical trials. *Journal of Clinical Oncology*, *32*(23), 2505–2511.

Jaki, T., Clive, S., & Weir, C. J. (2013). Principles of dose finding studies in cancer: A comparison of trial designs. *Cancer Chemotherapy and Pharmacology*, *71*(5), 1107–1114.

Ji, Y., Liu, P., Li, Y., & Nebiyou Bekele, B. (2010). A modified toxicity probability interval method for dose-finding trials. *Clinical Trials*, *7*(6), 235–244.

Lee, S. M., & Cheung, Y. K. (2009). Model calibration in the continual reassessment method. *Clinical Trials*, *6*(3), 227–238.

Lin, R., & Yin, G. (2015). Bayesian optimal interval design for dose finding in drug-combination trials. *Statistical Methods in Medical Research*, https://doi.org/10.1177/0962280215594494.

Liu, S., & Yuan, Y. (2015). Bayesian optimal interval designs for phase I clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *64*(3), 507–523.

Mander, A. P., & Sweeting, M. J. (2015). A product of independent beta probabilities dose escalation design for dual-agent phase I trials. *Statistics in Medicine*, *34*(8), 1261–1276.

O'Quigley, J., Pepe, M., & Fisher, L. (1990). Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics*, *46*(1), 33–48.

Riviere, M. K., Yuan, Y., Dubois, F., & Zohar, S. (2014). A Bayesian dose-finding design for drug combination clinical trials based on the logistic model. *Pharmaceutical Statistics*, *13*(4), 247–257.

Riviere, M. K., Yuan, Y., Dubois, F., & Zohar, S. (2015). A Bayesian dose-finding design for clinical trials combining a cytotoxic agent with a molecularly yargeted agent. *Journal of the Royal Statistical Society: Series C*, *64*, 215–229.

Rogatko, A., Schoeneck, D., Jonas, W., Tighiouart, M., Khuri, F. R., & Porter, A. (2007). Translation of innovative designs into phase I trials. *Journal of Clinical Oncology*, *25*(31), 4982–4986.

Simon, R., Rubinstein, L., Arbuck, S. G., Christian, M. C., Freidlin, B., & Collins, J. (1997). Accelerated titration designs for phase I clinical trials in oncology. *Journal of the National Cancer Institute*, *89*(15), 1138–1147.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, *45*(3), 925–937.

Storer, B. E. (2001). An evaluation of phase I clinical trial designs in the continuous dose-response setting. *Statistics in Medicine*, *20*(16), 2399–2408.

Stylianou, M., & Flournoy, N. (2002). Dose finding using the biased coin up-and-down design and isotonic regression. *Biometrics*, *58*(1), 171–177.

Thall, P. F., Millikan, R. E., Mueller, P., & Lee, S. J. (2003). Dose-finding with two agents in phase I oncology trials. *Biometrics*, *59*(3), 487–496.

van Brummelen, E. M. J., Huitema, A. D. R., van Werkhoven, E., Beijnen, J. H., & Schellens, J. H. M. (2016). The performance of model-based versus rule-based phase I clinical trials in oncology: A quantitative comparison of the performance of model-based versus rule-based phase i trials with molecularly targeted anticancer drugs over the last 2 years. *Journal of Pharmacokinetics and Pharmacodynamics*, *43*(3), 235–242.

Wages, N. A., Conaway, M. R., & O'Quigley, J. (2011). Continual reassessment method for partial ordering. *Biometrics*, *67*(4), 1555–1563.

Wang, K., & Ivanova, A. (2005). Two-dimensional dose finding in discrete dose space. *Biometrics*, *61*(1), 217–222.

Yan, F., Mandrekar, S. J., & Yuan, Y. (2017). Keyboard: A novel bayesian toxicity probability interval design for phase I clinical trials. *Clinical Cancer Research*, *23*(15), 3994–4003.

Yin, G., & Yuan, Y. (2009). Bayesian model averaging continual reassessment method in phase I clinical trials. *Journal of the American Statistical Association*, *104*(487), 954–968.

Yin, G., & Yuan, Y. (2009a). A latent contingency table approach to dose finding for combinations of two agents. *Biometrics*, *65*(3), 866–875.

Yin, G., & Yuan, Y. (2009b). Bayesian dose finding in oncology for drug combinations by copula regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *58*(2), 211–224.

Yuan, Y., Hess, K. R., Hilsenbeck, S. G., & Gilbert, M. R. (2016). Bayesian optimal interval design: A simple and well-performing design for phase I oncology trials. *Clinical Cancer Research*, *22*, 4291–4301.

Yuan, Y., Nguyen, H. Q., & Thall, P. F. (2016). *Bayesian Designs for Phase I–II Clinical Trials*. New York: Chapman & Hall/CRC.

Yuan, Y., & Yin, G. (2008). Sequential continual reassessment method for two-dimensional dose finding. *Statistics in Medicine*, *27*(27), 5664–5678.

Zhang, L., & Yuan, Y. (2016). A practical Bayesian design to identify the maximum tolerated dose contour for drug combination trials. *Statistics in Medicine*, *35*(27), 4924–4936.

Zhou, H., Murray, T. A., Pan, H., & Yuan, Y. (2018a). Comparative review of toxicity probability interval designs for phase I clinical trials. *Statistics in Medicine*, *37*(14), 2208–2222.

Zhou, H., Yuan, Y., and Nie L. (2018b). Accuracy, safety and reliability of novel Phase I trial designs. *Clinical Cancer Research*, https://doi.org/10.1158/1078-0432.CCR-18-0168.

Chapter 9
Data Monitoring: Structure for Clinical Trials and Sequential Monitoring Procedures

David M. Reboussin and Dave L. DeMets

Many aspects of developing and fielding a clinical trial can be guided by well-known theoretical properties of experimental design and practical experience with the conduct of other types of studies on human populations. Interim data monitoring, however, is often unfamiliar territory even to investigators who have been involved with other aspects of clinical trials. Yet there is some general agreement based on long experience, and some published literature, on successful models for structuring the monitoring process, for statistical methods that can address common questions relating to the decision to continue or stop a trial, and on how a committee should approach the decision in the context of what is often an unexpectedly complex set of issues. In this chapter we present a short discussion of these aspects of data monitoring in clinical trials. The fundamental need for monitoring in order to fulfil the investigators' ethical responsibilities to participants in a trial provides a rationale for the monitoring committee to be composed of experts independent of the investigators and suggests some basic structure for the committee's meetings. The nature of statistical assessment of accumulating data has produced a variety of methods which can address specific questions of importance for data monitoring and are flexible enough to accommodate some practical constraints under which the monitoring committee must act. Finally, the variety and complexity of actual data monitoring experiences shows that no statistical technique can be used as the sole basis in the decision to stop or continue: the monitoring committee's charge usually requires an active, thorough review of evidence which is neither simple nor straightforward.

D. M. Reboussin (✉) · D. L. DeMets
Department of Biostatistics, Wake Forest School of Medicine, Winston-Salem, NC, USA
e-mail: drebouss@wakehealth.edu

D. L. DeMets
e-mail: demets@biostat.wisc.edu

D. L. DeMets
Department Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_9

9.1 Monitoring Committee Structure and Function

Investigators in a clinical trial have an ethical responsibility to participants that demands both harms and benefits be monitored during trials. Early termination of the trial should be considered if data partway through the trial demonstrates harm attributable to the intervention, or if a clear benefit of the intervention creates ethical concerns for the control group. Alternatively, if differences in primary and possibly secondary response variables are so unimpressive that the prospect of a clear result is extremely unlikely, it may not be justifiable in terms of time, money, and effort to continue the trial. Monitoring of response variables can also identify the need to collect additional data to address questions of benefit or safety that arise during the trial. Finally, monitoring may reveal logistical or data quality problems that need to be addressed promptly. Thus, there are ethical, scientific, and economic reasons for interim evaluation of clinical trials (Baum et al. 1994; Fleming and DeMets 1993; Heart Special Project Committee 1988). The monitoring committee process has been described in detail (Ellenberg et al. 2003) as have case studies of trials terminated for benefit, harm, or futility (DeMets et al. 2006).

Keeping in mind these issues, data and safety monitoring is not simply a matter of looking at tables or results of statistical analysis of the primary outcome. Rather, it is an active process in which additional tabulations and analysis are suggested and evolve as a result of ongoing review. Monitoring also involves an interaction with the individuals responsible for collating, tabulating, and analyzing the data. For single center studies, the monitoring responsibility could, in principle, be assumed by the investigator. However, while monitoring the data, the investigator may discover trends toward benefit or harm while participants are still being enrolled or treated. Participants agree be part of clinical trials with the understanding that neither intervention nor control is favored, a state of clinical equipoise (Freedman 1987). Knowing that a trend exists may make it difficult for an investigator to continue to enroll, follow, evaluate, and care for the participants in an unbiased manner. The credibility of the trial is enhanced if, instead of the investigator, an independent person monitors the response variable data. Though some authors disagree (Crowley et al. 1994; Green and Crowley 1993; Harrington et al. 1994), these considerations suggest that individuals who monitor later phase clinical trials should have no formal involvement with the participants or the investigators.

One or two knowledgeable individuals may suffice for small, short-term studies whether early or late phase. For larger trials, the responsibility for monitoring response variable data is usually placed with an independent group acting as a committee (Ellenberg et al. 2003; DeMets et al. 2006; Fisher et al. 2001). Independence protects the members of the monitoring committee from being influenced in the decision-making process by investigators, participants and federal or industry sponsors. The committee would usually include experts in the relevant clinical fields or specialties, individuals with experience in the conduct of clinical trials, epidemiologists, biostatisticians knowledgeable in design and analysis, and often for NIH funded trials a bioethicist or participant advocate. While statistical procedures are

often helpful in evaluating interim results, the decision process to continue, terminate a trial early, or modify the design is invariably complex and no single statistical procedure is adequate to address all trials.

The first priority of the monitoring committee must be to ensure the safety of participants in the trial. The second priority is meeting obligations to the investigators and the Institutional Review Boards or ethics committees, who place enormous trust in the monitoring committee both to protect participants from harm and ensure the integrity of the trial. Third, the monitoring committee has a responsibility to the sponsor of the trial, whether federal or private. Finally, the monitoring committee provides a service to drug or device regulatory agencies, especially for trials which are utilizing drugs, biologics or devices which still have investigational status.

Although many formats for monitoring committee meetings have been used, a format that includes an open session, a closed session, and an executive session allows for exchange of information by all relevant parties and for appropriate confidential and independent review (Ellenberg et al. 2003; DeMets et al. 1995). The open session enables interaction between investigator representatives such as the study principal investigator or chair, the sponsor, the statistical center, the industrial partners, and the monitoring committee. Though uncommon, it may sometimes be appropriate for a regulatory agency to participate in the open session. In this session, issues of participant recruitment, data quality, general adherence, toxicity issues, and any other logistical matter that may affect either the conduct or outcome of the trial are considered in a blinded fashion. In a closed session with monitoring committee members and one or more members of the statistical team, analyses of the confidential unblinded outcome data are reviewed. This review includes comparison by intervention groups of baseline variables, primary or secondary variables, safety or adverse outcome variables, adherence measures for the entire group, and examinations of any relevant subgroups. Following this review, the monitoring committee may decide to have an executive session with its members only where decisions about continuation, termination or protocol modification are made. After the closed session, they may meet with a representative of the sponsor or investigator leadership to share their recommendations which are usually followed up in a letter. Regardless of how formal, most monitoring committee meetings have these components.

Who specifically attends the various sessions must be decided before the trial begins and before the first monitoring committee meeting is scheduled. In general, attendance should be limited to those who are essential for proper monitoring. As noted, it is common for the study principal investigator and sponsor representatives to attend the first open session. If he or she does not provide care for participants in the trial, the principal investigator will sometimes attend the closed session; however, that practice is not recommended. If the study is sponsored by industry, independence and credibility of the study is best served by no industry attendance at the closed session. Industry sponsored trials that are also managed and analyzed by industry will require a statistician from the sponsor who prepares the monitoring report to attend. In such situations the company statistician must have a "firewall" separating her from other colleagues at the company, something that may be difficult to achieve in a way that is convincing to outsiders. However, another common practice for industry-sponsored

pivotal Phase III trials is for a separate statistical analysis center to provide the interim analyses and report to the independent monitoring committee (Fisher et al. 2001). This practice reduces the possibility or perception that interim results are known to the industry sponsor or the investigator group. Regulatory agency representatives usually do not attend the closed session because being involved in the monitoring decision may affect their regulatory role should the product be subsequently submitted for approval. An executive session should involve only the voting members of the monitoring committee, although an independent statistician who provided the data report may also attend.

How the intervention or treatment comparisons will be presented to the monitoring committee also must be resolved before the start of a trial. In some trials, the monitoring committee knows the identity of the interventions in each table or figure of the report. In other trials, for two interventions the tables may be labelled as A and B with the identity of A and B remaining blinded until the monitoring committee requests the unblinding on a "need to know" basis. Thus, if there are no trends in either benefit or harm, which is likely to be the case early in a trial, there is no overwhelming reason to know the identity of groups A and B. When trends begin to emerge in either direction, the monitoring committee should have full knowledge of the group identities (Meinert 1998). In some trials, the monitoring committee is blinded throughout the interim monitoring. While this degree of blinding may enhance objectivity, it conflicts with the monitoring committee's primary purpose of protecting the participants in the trial from harm or unnecessary continuation. As pointed out by Whitehead (1999), the intention of this approach is to deny the monitoring committee a complete picture of the interim data. To assess the progress of the trial, the harm and benefit profile of the intervention must be well understood and the possible tradeoffs weighed. If each group of tables is labeled by a different code, the committee cannot easily assess the overall harm/benefit profile of the intervention, and thus may put participants at unnecessary risk or continue a trial beyond the point at which benefit outweighs risks. Such complex coding schemes also increase the chance for errors in labeling. This practice is not common and not recommended.

9.2 Statistical Methods Used in Interim Monitoring

The previous section discussed the administrative structure was for conducting interim analysis of data quality and outcome data for benefit and potential harm to trial participants. This section reviews some statistical methods for sequential analysis that are currently used for monitoring accumulating data in a clinical trial. These methods help inform decisions as to whether the trial should be terminated early for benefit, harm, or futility or whether it should be continued to its planned termination. No single statistical test or monitoring procedure should be used as a strict rule for decision-making, but rather as one piece of evidence to be integrated with the totality of evidence (Fleming and DeMets 1993; Ellenberg et al. 2003; DeMets et al. 2006; Fisher et al. 2001; Canner 1981; DeMets 1990). Therefore, it is difficult

to make a single recommendation about which statistical approach should be used. However, the following methods, when applied appropriately, can be useful guides in the decision-making process.

Repeated Testing for Significance

Repeated significance testing of accumulating data, which is essential to monitoring, has statistical implications (The Coronary Drug Project Research Group 1975; Armitage et al. 1969; Armitage 1957; Bross 1952; Anscombe 1963; Robbins 1952, 1970). If the null hypothesis of no difference between two groups is true, and tests of that hypothesis are made repeatedly at the same level of significance using accumulating data, the probability that the test will be declared significant by chance alone is larger than that significance level. That is, the rate of incorrectly rejecting the null hypothesis, a false positive error, will be larger than if only a single test had been done. Regardless of the test statistic or type of outcome, repeated testing of accumulating data without taking into account the number of tests increases the overall probability of incorrectly rejecting the null hypothesis H_0 and claiming an intervention effect. If the repeated testing continues indefinitely, the null hypothesis is certain to be rejected eventually, but even five or ten tests can lead to a misinterpretation of the results of a trial if the multiple testing issues are ignored.

As an example, in a clinical trial where the participant response is known relatively soon after entry the difference in rates between two groups may be compared repeatedly as participants are recruited. The usual test statistic for comparing two proportions is the chi-square test or the equivalent normal test statistic. The null hypothesis is that the true response rates or proportions are equal. If a significance level of 5% is selected and the null hypothesis, H_0, is tested only once, the probability of rejecting H_0 if it is true is 5% by definition. However, if H_0 is tested twice, first when one-half of the data are known and then when all the data are available, the probability of incorrectly rejecting H_0 is increased from 5 to 8% (Armitage et al. 1969). If the hypothesis is tested five times, with one-fifth of the participants added between tests, the probability of finding a significant result if the usual statistic for the 5% significance level is used becomes 14%. For ten tests, this probability increases to almost 20%. The methods below describe ways to adjust for this inflation of type 1 error that were developed initially for means and proportions, but they can be validly applied to most common trial designs. Under very general conditions, any statistic used for repeated testing of a single parameter from a parametric or semiparametric statistical model has a normal or asymptotically normal distribution with a known correlation structure over time (Jennison and Turnbull 1997; Scharfstein et al. 1997). Besides logrank and other survival tests, comparisons of means, comparison of proportions (Kim and DeMets 1992; Pocock 1977) and comparison of linear regression slopes (Lee 1994; Lee and DeMets 1991, 1992; Su and Lachin 1992; Wei et al. 1990; Wu and Gordon Lan 1992) can be monitored with the same methods.

A classic illustration of the repeated testing problem is provided by the Coronary Drug Project (CDP) for the clofibrate versus placebo mortality comparison, shown in Fig. 9.1 (Canner 1981; The Coronary Drug Project Research Group 1975). The CDP was a long-term randomized, double-blind, multicenter study that compared the

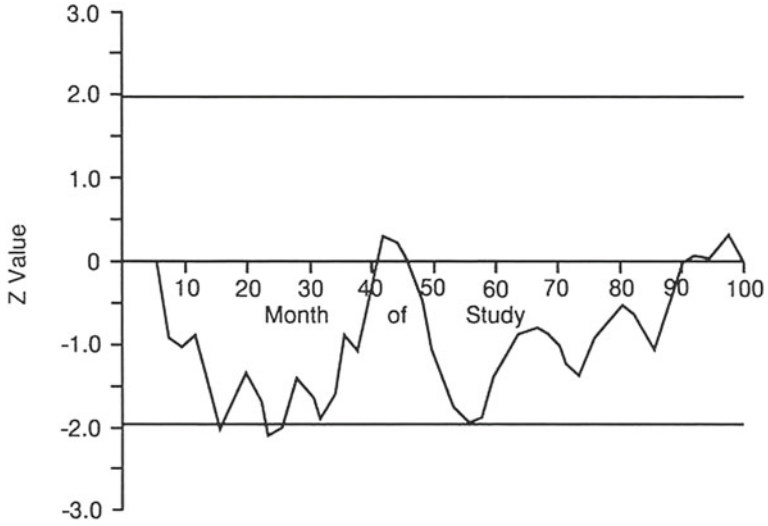


Fig. 9.1 Interim survival analyses comparing mortality in clofibrate- and placebo-treated participants in the Coronary Drug Project (Canner 1981). A positive Z value favors placebo

effect on total mortality of several lipid-lowering drugs (high- and low-dose estrogen, dextrothyroxine, clofibrate, nicotinic acid) against placebo. Figure 9.1 presents the standardized mortality comparisons over the follow-up or calendar time of the trial. The two horizontal lines indicate the conventional value of the test statistic, corresponding to a two-sided 0.05 significance level, used to judge statistical significance for studies where the comparison is made just one time. It is evident that the trends in this comparison emerge and weaken throughout, coming close or exceeding the conventional critical values on five monitoring occasions. However, as shown in Fig. 9.2, the mortality curves at the end of the trial are nearly identical, corresponding to the very small standardized statistic at the end of the Fig. 9.1. The monitoring committee for this trial took into consideration the repeated testing problem and did not terminate this trial early just because the conventional significance values were exceeded.

Group Sequential Methods

Classical sequential methods were developed to minimize the number of participants required for a study. Continued enrollment of participants depends on results from those already entered. Most of these sequential methods assume that the response variable outcome is known in a short time relative to the duration of the trial, as is true for many trials involving acute illness. For studies involving chronic diseases, classical sequential methods have not been as useful. Detailed discussions of classical sequential methods are given, for example, by Armitage (1975), Whitehead (1997), and Wald (2013).

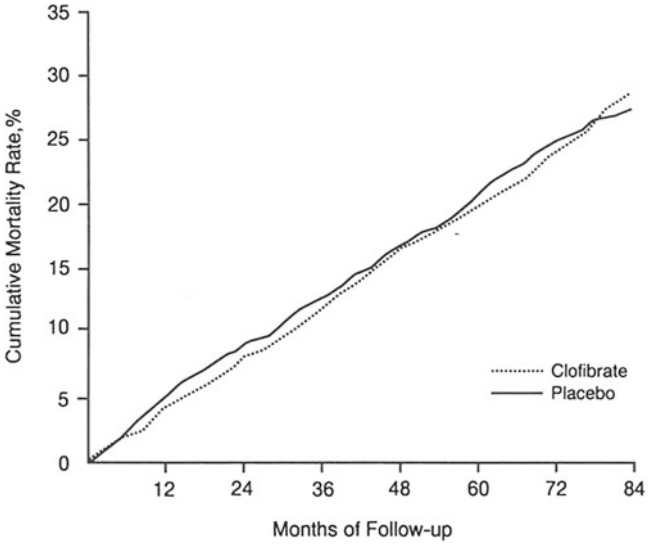


Fig. 9.2 Cumulative mortality curves comparing clofibrate- and placebo treated participants in the Coronary Drug Project (Canner 1981)

The limitations of classical sequential methods led to the development of other approaches to the repeated testing problem. Ad hoc rules have been suggested that attempt to ensure a conservative interpretation of interim results. One such method is to use a critical value of 2.6 for the normal test statistic Z-value at each interim look as well as in the final analyses (Canner 1981). Another approach (Haybittle 1971; Peto et al. 1976) referred to as the Haybittle–Peto procedure, favors using a large critical value such as 3.0 for all interim tests. With this procedure any adjustment needed for repeated testing at the final test is negligible and the conventional critical value can be used. These methods are ad hoc in the sense that no precise Type I error level is guaranteed.

Pocock (1977, 1978, 1982) modified the repeated testing methods of McPherson and Armitage (1971) and developed a group sequential method for clinical trials which avoids many of the limitations of classical methods and guarantees a pre-specified type 1 error level. He discusses two cases of special interest; one for comparing two proportions and another for comparing mean levels of response. Pocock's method divides the participants into a series of K equal-sized groups with $2n$ participants in each, n assigned to intervention and n to control. K is the number of times the data will be monitored during the course of the trial. The total expected sample size is $2nK$. The test statistic used to compare control and intervention is computed as soon as data for the first group of $2n$ participants are available, and recomputed when data from each successive group of $2n$ participants become known. Under the null hypothesis, the distribution of the test statistic Z_i is assumed to be approximately normal with zero mean and unit variance, where i indicates the group number ($i \leq$

K). This statistic is compared to the stopping boundaries, $\pm ZN_K$ where ZN_K has been determined to assure that the overall (two sided) significance level for the trial will be α when up to K repeated tests are done. For example, if $K=5$ and $\alpha=0.05$ (two-sided), $ZN_K=2.413$. This critical value is larger than the critical value of 1.96 used in a single test of hypothesis with $\alpha=0.05$. If the statistic Z_i falls outside the boundaries on the "i"-th repeated test, indicating rejection of the null hypothesis and suggesting that the trial should be terminated. If the statistic never falls outside the boundaries, the trial should be continued until $i=K$ (the maximum number of tests). When $i=K$, the trial would stop and the investigator would "accept" H_0. O'Brien and Fleming (1979) discuss a similar group sequential procedure. Using the above notation, their stopping rule compares the statistic Z_i with $Z^* \sqrt{(K/i)}$ where Z^* is determined so as to achieve the desired significance level. For example, if $K=5$ and $a=0.05$, $Z^*=2.04$. If $K \leq 5$, Z^* may be approximated by the usual critical values for the normal distribution. One attractive feature is that the critical value used at the last test ($i=K$) is approximately the same as that used if a single test were done.

These group sequential methods have an advantage over the classical methods in that the data do not have to be continuously tested and individual participants do not have to be enrolled in pairs. This concept suits the data review activity of most large clinical trials where monitoring committees meet periodically. Furthermore, in many trials continuous consideration of early stopping is unnecessary. Pocock (1977, 1978, 1982) discusses the benefits of the group sequential approach in more detail and other authors describe variations (DeMets 1987; Fleming 1990; Fleming and Watelet 1989; Freedman et al. 1983; Jennison and Turnbull 1990).

In Fig. 9.3 boundaries for the Haybittle-Peto, Pocock and O'Brien-Fleming methods described are given for $K=5$ and $\alpha=0.05$ (two-sided). If for $i<5$ the test statistic falls outside the boundaries, the trial is terminated and the null hypothesis rejected. Otherwise, the trial is continued until $i=5$, at which time the null hypothesis is either rejected or "accepted". The three boundaries have different early stopping properties. The O'Brien–Fleming model is unlikely to lead to stopping in the early stages. Later on, however, this procedure leads to a greater chance of stopping prior to the end of the study than the other two. Both the Haybittle–Peto and the O'Brien–Fleming boundaries avoid the awkward situation of accepting the null hypothesis when the observed statistic at the end of the trial is much larger than the conventional critical value (i.e., 1.96 for a two-sided 5% significance level). If the observed statistic in Fig. 9.3 is 2.3 when $i=5$ the result would not be significant using the Pocock boundary. The large critical values used at the first few analyses for the O'Brien–Fleming boundary can be adjusted to some less extreme values (e.g. 3.5) without noticeably changing the interim and final critical values used later.

Many monitoring committees often wish to be somewhat conservative in their interpretation of early results because of the uncertainties discussed earlier and because a few additional events can alter the results substantially. Yet, most investigators would like to use conventional critical values in the final analyses, not requiring any penalty for interim analyses. This means that the critical value used in a conventional fixed sample methods would be the same for that used in a sequential plan, resulting in no increase in sample size. With that in mind, the O'Brien–Fleming pro-

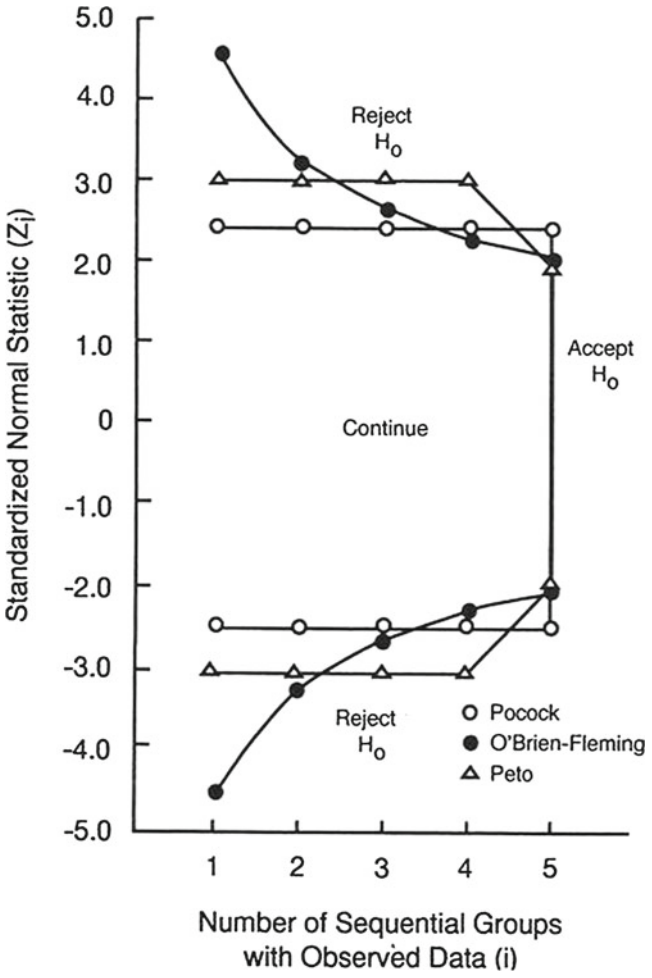


Fig. 9.3 Three group sequential stopping boundaries for the standardized normal statistic (Z_i) for up to five sequential groups with two-sided significance level of 0.05 (DeMets and Lan 1994)

cedure has considerable appeal, perhaps with the adjusted or modified boundary as described.

The Beta-Blocker Heart Attack Trial (BHAT) provides an example where sequential monitoring led to early termination (DeMets et al. 1984; Beta-Blocker Heart Attack Trial Research Group 1982). This randomized placebo control trial enrolled over 3800 participants with a recent myocardial infarction to evaluate the effectiveness of propranolol in reducing mortality. Interim log-rank tests were evaluated using the O'Brien–Fleming group sequential procedure (O'Brien and Fleming 1979). Seven meetings had been scheduled to review interim data. The trial was designed for a two-sided 5% significance level. These specifications produce the group sequential

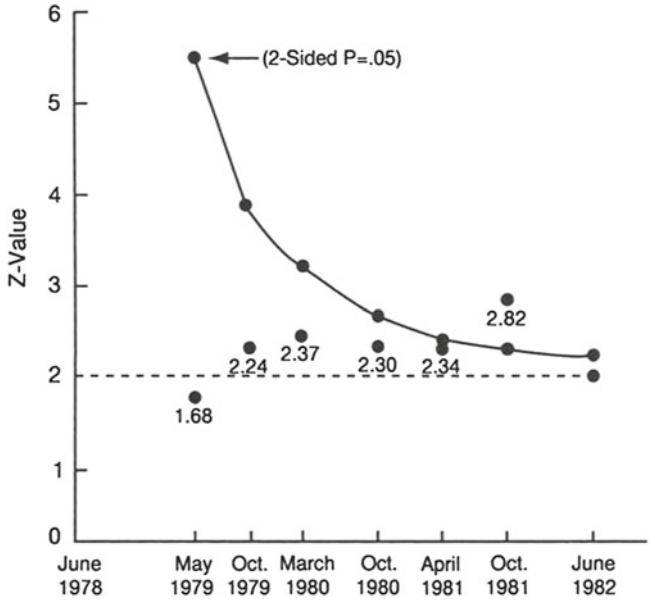


Fig. 9.4 Six interim log rank statistics plotted for the time of data monitoring committee meetings with a two-sided O'Brien-Fleming significance level boundary in the Beta-Blocker Heart Attack Trial (DeMets et al. 1984). Dashed line represents $Z = 1.96$

boundary shown in Fig. 9.4. In addition, the interim results of the log-rank statistic are also shown for the first six meetings. From the second analysis on, the conventional significance value of 1.96 was exceeded. Nevertheless, the trial was continued. At the sixth meeting, after an average of a little over 2 years of a planned 3 year follow-up, a mortality difference was observed, and the O'Brien–Fleming boundary was crossed as shown in Fig. 9.4. The results were statistically significant, allowing for repeated testing, and with high probability would, not be reversed during the next year (DeMets et al. 1984). However, it should be emphasized that crossing the boundary was not the only factor in this decision. The data monitoring committee debated whether the additional year of follow-up would add valuable information. It was argued that there would be too few events in the last year of the trial to provide a good estimate of the effect of propranolol treatment in the third and fourth year of therapy. Thus, the committee decided that prompt publication of the observed benefit was more important than waiting for the marginal information yet to be obtained. This trial was one of the early trials to implement group sequential monitoring boundaries discussed below and will be used as an example to illustrate the method.

Flexible Group Sequential Procedures: Alpha Spending Functions

While the group sequential methods described above are an important advance in data monitoring, the Beta-blocker Heart Attack Trial (BHAT) (DeMets et al. 1984; Beta-Blocker Heart Attack Trial Research Group 1982) experience suggested two

limitations. One was the need to specify the number K of planned interim analyses in advance. The second was the requirement for equal numbers of either participants or events between each analysis. In the BHAT example, the numbers of deaths between analyses were not equal and exactly seven analyses of the data had been specified. If the monitoring committee had requested an additional analysis between the fifth and sixth scheduled meetings, the O'Brien–Fleming group sequential procedure would not have directly accommodated such a modification. Yet such a request could easily have happened. In order to accommodate the unequal numbers of participants or events between analyses and the possibility of larger or fewer numbers of interim analyses than pre-specified, flexible procedures eliminating those restrictions were developed (DeMets and Lan 1994; Kim and DeMets 1987; Lan and DeMets 1989; Lan and DeMets 1983; Lan et al. 1984, 1993, 1994; Lan and Zucker 1993). The authors proposed a so-called alpha spending function which allows investigators to determine how they want to allocate or "spend" the Type I error or alpha during the course of the trial. This function guarantees that at the end of the trial, the overall Type I error will equal the prespecified value of α. As will be described, this approach is a generalization of the previous group sequential methods so that the Pocock (1977) and O'Brien and Fleming (1979) monitoring procedures become special cases.

To understand how this flexibility is incorporated, we must distinguish between calendar time and information fraction (Lan and DeMets 1989; Lan et al. 1994). The information expected from all participants at the planned end of the trial is the total information. At any particular calendar time t during the study, a certain fraction t^* of the total information is observed. The fraction of total information may be approximated as a fraction of participants randomized at that point (n randomized divided by the total number of expected randomizations, N) or in survival studies by a fraction of observed events (d observed events divided by the total number of expected events, D). Thus the value for t^* must be between 0 and 1. The information fraction is more generally defined in terms of ratio of the inverse of the variance of the test statistic at the particular interim analysis and the final analysis. The alpha spending function, $\alpha(t^*)$, determines how the prespecified α is allocated at each interim analyses as a function of the information fraction. At the beginning of a trial, $t^* = 0$ and $\alpha(t^*) = 0$, while at the end of the trial, $t^* = 1$ and $\alpha(t^*) = \alpha$. Alpha-spending functions that correspond to the Pocock and O'Brien–Fleming boundaries in Fig. 9.3 are shown in Fig. 9.5 for a two-sided 0.05 α level and five interim analyses. These spending functions correspond to interim analyses at information fractions at 0.2, 0.4, 0.6, 0.8, and 1.0. However, in practice the information fractions need not be equally spaced.

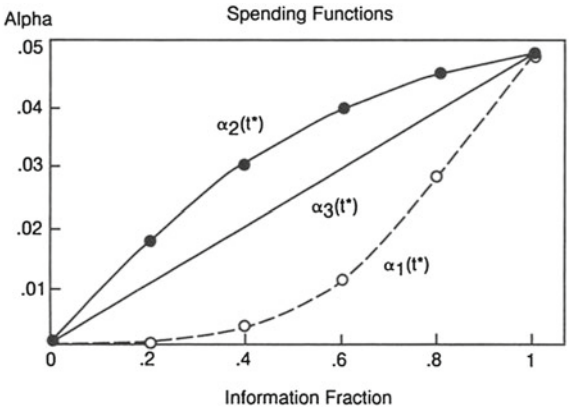


Fig. 9.5 Alpha-spending functions for $K = 5$ with two-sided $\alpha = 0.05$ at information fractions 0.2, 0.4, 0.6, 0.8, and 1 (DeMets and Lan 1995). $\alpha_1(t^*) \sim$ O'Brien-Fleming; $\alpha_2(t^*) \sim$ Pocock; $\alpha_3(t^*) \sim$ uniform

Many different spending functions can be specified. The O'Brien–Fleming $\alpha_1(t^*)$, Pocock $\alpha_2(t^*)$ and uniform $\alpha_3(t^*)$ type spending functions are specified as follows:

O'Brien-Fleming	$\alpha_1(t^*) = 2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{t^*}\right)$
Pocock	$\alpha_2(t^*) = \alpha \ \ln(1 + (e-1)t^*)$
Uniform	$\alpha_3(t^*) = \alpha\, t^{*\theta}$ for $\theta > 0$

The spending function $\alpha_3(t^*)$ spends alpha uniformly during the trial for $\theta = 1$, at a rate between $\alpha_1(t^*)$ and $\alpha_2(t^*)$. Other spending functions have also been defined (Hwang et al. 1990; Wang and Tsiatis 1987). The Pocock-type spending function allocates a greater proportion of the alpha earlier than the O'Brien–Fleming type spending function. For the O'Brien–Fleming-type spending function at $t^* = 0.2$, $\alpha(0.2)$ is less than 0.0001 which corresponds approximately to the very large critical value or boundary value of 4.56 in Fig. 9.3. At $t^* = 0.4$, the amount of α which can be spent is $\alpha(0.4) - \alpha(0.2)$ which is approximately 0.0006, corresponding to the boundary value 3.23 in Fig. 9.3. Obtaining these critical values requires numerical integration and is described elsewhere in detail (Lan and DeMets 1983). Programs are available for these calculations (Reboussin et al. 2000, 2003).

The advantage of the alpha-spending function is that neither the number nor the time of the interim analyses needs to be specified in advance. Once the particular spending function is selected, the information fractions t_1^*, t_2^*, … determine the critical or boundary values exactly. In addition, the frequency of the interim analyses can be changed during the trial and still preserve the prespecified α level. Even if the rationale for changing the frequency is dependent on the emerging trends, the impact on the overall Type I error rate is almost negligible (Lan and DeMets 1989; Proschan et al. 1992). These advantages give the spending function approach to group sequential monitoring the flexibility in analysis times that is often required in actual clinical trial settings (Geller 1994). It must be emphasized that no change of the

spending function itself is permitted during the trial. Other authors have discussed further aspects of this approach (Falissard and Lellouch 1992; Lan and Lachin 1990; Li and Geller 1991).

p-values and Confidence Intervals for Group Sequential Procedures

If the trial continues to the scheduled termination point, a p value is often computed to indicate the extremeness of the result. If the standardized statistical test exceeds the critical value, the p value would be less than the corresponding significance level (e.g. $p < 0.05$). If a trial is terminated early or continues to the end with the standardized test exceeding or crossing the boundary value, a p value can also be computed (Gange and DeMets 1996). These p values must account for the repeated statistical testing of the outcome measure and for the particular monitoring boundary employed. Calculation of the p value is relatively straightforward with existing software packages (Reboussin et al. 2000, 2003).

Statistical tests of hypotheses are but one of the methods used to evaluate the results of a clinical trial. Once trials are terminated, either on schedule or earlier, confidence intervals (CIs) are often used to give some sense of the uncertainty in the estimated treatment or intervention effect. For a fixed sample study, CIs are typically constructed as

$$(\text{effect estimate}) \pm Z(\alpha)\ \text{SE}(\text{estimate})$$

where SE is the standard error of the estimate. In the group sequential monitoring setting, this CI is referred to as the naïve estimate since it does not take into account the sequential testing aspects. In general, construction of CIs following the termination of a clinical trial is not as straightforward (Chang and O'Brien 1986; DeMets and Lan 1989; Emerson and Fleming 1990; Hughes and Pocock 1988; Jennison and Turnbull 1984, 1989; Kim 1989; Kim and DeMets 1987; Pocock and Hughes 1989; Rosner and Tsiatis 1988; Siegmund 1978; Tsiatis et al. 1984; Whitehead 1986; Whitehead and Facey 1991), but software exists to aid in the computations (Reboussin et al. 2000). The major problem with naive CIs is that they may not give proper coverage of the unknown but estimated treatment effect; that is, they may not include the true effect with the specified frequency (e.g. 95%). To construct a better interval estimate, a rule to determine which of two test statistics at different times is more extreme. Such a rule orders the possible sequential outcomes for a trial, and several different ordering rules have been proposed (Chang and O'Brien 1986; DeMets and Lan 1989; Emerson and Fleming 1990; Hughes and Pocock 1988; Jennison and Turnbull 1984, 1989; Kim 1989; Kim and DeMets 1987; Pocock and Hughes 1989; Rosner and Tsiatis 1988; Siegmund 1978; Tsiatis et al. 1984; Whitehead 1986; Whitehead and Facey 1991). None of the rules proposed appear to be universally superior but the ordering originally suggested by Siegmund (1978) and adopted by Tsiatis et al. (1984) is quite adequate in most circumstances. In this ordering, any treatment comparison statistic which exceeds the group sequential boundary at one time is considered to be more extreme than any result which exceeds the sequential boundary at a later time. While construction of CIs using this ordering of possible outcomes can break

down in certain cases, such cases are quite unusual and not likely to occur in practice (Whitehead and Facey 1991). It is also interesting that for conservative monitoring boundaries such as the O'Brien–Fleming method, the naive CI does not perform that poorly, due primarily to the extreme early conservatism of the boundary (Rosner and Tsiatis 1988). While more exact CIs can be computed for this case, the naive estimate may still prove useful as a quick estimate to be recalculated later (Tsiatis et al. 1984). Pocock and Hughes (1989) have suggested that the point estimate of the effect of the intervention should also be adjusted, since trials that are terminated early tend to exaggerate the size of the true treatment difference. Others have also pointed out the bias in the point estimate Emerson and Fleming (1990; Kim 1989). Kim (1989) suggested that an estimate of the median effect is less biased.

Asymmetric Boundaries

In most trials, the main purpose is to test whether the intervention is superior to the control. It is rarely ethical to continue a study in order to prove, at the usual levels of significance, that the intervention is harmful relative to a placebo or standard control. This point has been mentioned by authors (DeMets and Ware 1980; DeMets and Ware 1982) who discuss methods for group sequential designs in which the hypothesis to be tested is one-sided; that is, to test whether the intervention is superior to the control. They proposed retaining the group sequential upper boundaries of methods such as Pocock, Haybittle–Peto, or O'Brien–Fleming for rejection of H_0 while suggesting various forms of a lower boundary which would imply "acceptance" of H_0. One simple approach is to set the lower boundary at an arbitrary value of Z_i such as -1.5 or -2.0. If the test statistic goes below that value, the data may be sufficiently suggestive of a harmful effect to justify terminating the trial. This asymmetric boundary attempts to reflect the behavior or attitude of members of many monitoring committees, who recommend stopping a study once the intervention shows a strong but non-significant trend in an adverse direction for major events. Emerson and Fleming (1989) recommend a lower boundary for acceptance of the null hypothesis which allows the upper boundary to be changed in order to preserve the Type I error exactly. Work by Gould and Pecore (1982) suggests ways for early acceptance of the null hypothesis while incorporating costs as well. For new interventions, trials might well be terminated when the chances of a positive or beneficial result seem remote (see below). However, if the intervention being is already in widespread use, it may be important to distinguish between lack of benefit and harm (DeMets et al. 1999). For example, if the intervention is not useful for the primary outcome, but not harmful, it may still have benefits such as on other secondary clinical outcomes, quality of life, or reduction in adverse events that would still make it a therapeutic option. In such cases, a symmetric boundary for the primary outcome might be appropriate.

An example of asymmetric group sequential boundaries is provided by the Cardiac Arrhythmia Suppression Trial (CAST). Two arms of the trial (encainide and flecainide, each vs. placebo) were terminated early using a symmetric two-sided boundary, although the lower boundary for harm was described as advisory by the authors (Friedman et al. 1993; Trial and Investigators 1992; Pawitan and Hallstrom 1990). The third comparison (moricizine vs. placebo) continued. However, due to

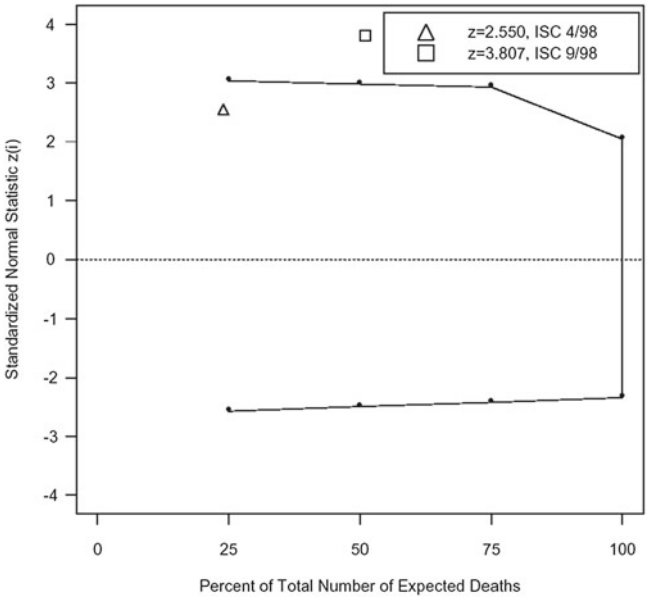


Fig. 9.6 MERIT-HF Group Sequential Monitoring Bounds for Mortality (Feyzi et al. 2006)

the experience with the encainide and flecainide arms, the lower boundary for harm was revised to be less stringent than originally, i.e. an asymmetric boundary was used (Trial and Investigators 1992).

MERIT-HF used a modified version of the Haybittle–Peto boundary for benefit, requiring a critical value near +3.0 and a similar but asymmetric boundary, close to a critical Z value of -2.5 for harm as shown in Fig. 9.6. In addition, at least 50% of the designed person years of exposure were to be observed before early termination could be recommended. The planned interim analyses to consider benefit were at 25, 50, and 75% of the expected target number of events. Because there was a concern that treating heart failure with a beta blocker might be harmful, the monitoring committee was required to evaluate safety on a monthly basis using the lower sequential boundary as a guide. At the 25% interim analyses, the statistic for the logrank test was +2.8, just short of the boundary for benefit. At the 50% interim analyses, the observed logrank statistic was +3.8, clearly exceeding the sequential boundary for benefit. It also met the desired number of person years of exposure as plotted in Fig. 9.6. Details of this experience are described elsewhere (Feyzi et al. 2006). A more detailed presentation of group sequential methods for interim analysis of clinical trials may be found in books by Jennison and Turnbull (Jennison and Turnbull 1999) and Proschan et al. (2006).

Curtailed Sampling and Conditional Power Procedures

During the course of monitoring accumulating data, one question often posed is whether the current trend in the data is so impressive that "acceptance" or rejection

of H_0 is already determined, or at least close to being determined. If the results of the trial are such that the conclusions are known for certain, no matter what the future outcomes might be, then consideration of early termination is in order. A helpful sports analogy is a baseball team "clinching the pennant" after winning a specific game: at that time, it is known for certain who has won and who has not won the pennant or league championship, regardless of the outcome of the remaining games. Playing the remaining games is done for reasons (e.g., fiscal) other than deciding the winner. This idea has been developed for clinical trials and is often referred to as deterministic curtailed sampling. It should be noted that group sequential methods focus on existing data while curtailed sampling in addition considers the data which have not yet been observed.

In some clinical trials, the final outcome may not be absolutely certain, but almost so. To use the baseball analogy again, a first place team may not have clinched the pennant but be so many games in front of the second place team that it is highly unlikely that it will not, in fact, end up the winner. Another team may be so far behind that it cannot "realistically" catch up. In clinical trials, this idea is often referred to as stochastic curtailed sampling or conditional power. Unconditional power is the probability at the beginning of the trial of achieving a statistically significant result at a prespecified alpha level and with a prespecified alternative treatment effect. Ideally, trials should be designed with a power of 0.80–0.90 or higher. However, once data begin to accumulate, the probability of attaining a significant result increases or decreases with emerging positive or negative trends. Calculating the probability of rejecting the null hypothesis of no effect once some data are available is conditional power.

Lan et al. (1982) considered the effect of stochastic curtailed or conditional power procedures on Type I and Type II error rates. If the null hypothesis H_0 is tested at time t using a statistic $S(t)$ then at the scheduled end of a trial at time T, the statistic would be denoted $S(T)$. Two cases are considered. First, suppose a trend in favor of rejecting H_0 is observed at time $t<T$, with intervention doing better than control. One then computes the conditional probability, γ_0 of rejecting H_0 at time T; that is, $S(T)$ greater than the final critical value assuming H_0 to be true and given the current value of $S(t)$. If this probability is sufficiently large, one might argue that the favorable trend is not going to disappear. Second, suppose a negative trend or data consistent with the null hypothesis of no difference, at some point t. Then, one computes the conditional probability $\gamma_1$of rejecting H_0 at the end of the trial given that some alternative H_1 is true. This addresses how large the true effect must be to reverse the observed "negative" trend. If the probability of a trend reversal is very low for a realistic range of alternative hypotheses, trial termination might be considered.

Because there is a small probability that future data may reverse an interim trend, a slightly greater risk of a Type I or Type II error will exist than would be if the trial continued to the scheduled end (Halperin et al. 1982). However, it has been shown that the Type I error is bounded very conservatively by α/γ_0 and the Type II error by β/γ_1. For example, if the probability of rejecting the null hypothesis, given the existing data were 0.85, then the actual Type I error would be no more than 0.05/0.85 or 0.059, instead of 0.05. The actual upper limit is considerably closer to 0.05,

Table 9.1 Expressions for the intervention effect θ under different outcomes

Outcome	Alternative θ	Notes
Survival	$\theta = \sqrt{D/4}\log(\lambda_C/\lambda_T)$	D=total events λ_C and λ_T are the hazard rates in the control and intervention arms, respectively
Binomial	$\theta = \dfrac{P_C - P_T}{\sqrt{2\bar{p}(1-\bar{p})/(n/2)}}$ $= \dfrac{(P_C - P_T)\sqrt{N/4}}{\sqrt{\bar{p}(1-\bar{p})}}$ $= 1/2\dfrac{(P_C - P_T)\sqrt{N}}{\sqrt{\bar{p}\bar{q}}}$	$2n = N$, n/arm or N = total sample size P_C and P_T are the event rates in the control arm and intervention arm respectively and $\bar{p}$ is the common event rate
Continuous (means)	$\theta = \left(\dfrac{\mu_C - \mu_T}{\sigma}\right)\sqrt{N/4}$ $= 1/2\left(\dfrac{\mu_C - \mu_T}{\sigma}\right)\sqrt{N}$	N=total sample size μ_C and μ_T are the mean response levels for the control and the intervention arms, respectively, and σ is the common standard deviation

but that calculation requires computer simulation. Calculation of these probabilities is relatively straightforward and the details have been described by Lan and Wittes (1988). A summary of these methods, using the approach of DeMets (2006), follows.

Let $Z(t)$ represent the standardized statistic or Z value at information fraction t. The conditional power (CP) for some alternative intervention effect θ, using a critical value of Z_α for a Type I error of alpha, can be calculated as

$$P[Z(1) \geq Z_\alpha | Z(t), \theta] = 1 - \Phi\left\{\left|Z_\alpha - Z(t)\sqrt{t} - \theta(1-t)\right|/\sqrt{1-t}\right\}$$

where $\theta = E(Z(t=1))$, the expected value of the test statistic at the full completion of the trial. The alternative θ is defined for various outcomes in Table 9.1.

If a particular value of the conditional power γ is selected as a "futility cutoff" then a boundary can also be produced which would indicate that if the test statistic fell below that, the chance of finding a significant result at the end of the trial is less than γ (Halperin et al. 1982). For example, in Fig. 9.7 the lower futility boundary is based on a specified conditional power γ, ranging from 10 to 30% that might be used to claim futility of finding a positive beneficial claim at the end of the trial. If the standardized statistic crosses that 20% lower boundary, the conditional power for a beneficial result at the end of the trial is less than 0.20 for the specified alternative.

Conditional power calculations are done for a specific alternative but in practice, a monitoring committee would likely consider a range of possibilities. These specified

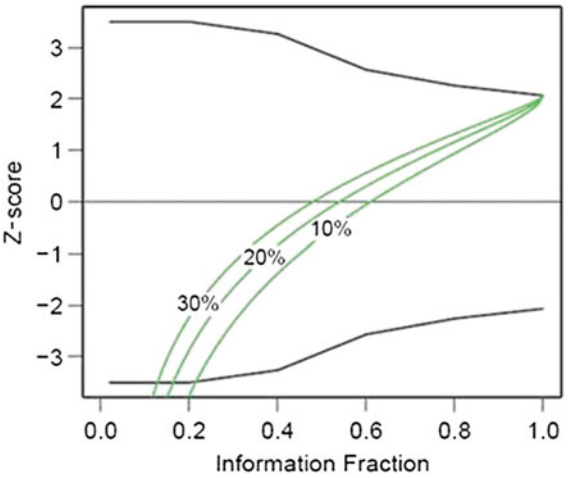


Fig. 9.7 Conditional Power Boundaries: outer boundaries represent symmetric O'Brien-Fleming type sequential boundaries ($\alpha = 0.05$). Three lower boundaries represent boundaries for 10, 20 and 30% conditional power to achieve a significant ($p < 0.05$) result of the trial conclusion (DeMets 2006)

alternatives may range between the null hypothesis of no effect and the prespecified design based alternative treatment effect. In some cases, a monitoring committee may consider even more extreme beneficial effects to determine just how much more effective the treatment would have to be to raise the conditional power to desired levels. These conditional power results can be summarized in a table or a graph, and then monitoring committee members can assess whether they believe recovery from a substantial negative trend is likely.

One of the earliest applications of conditional power was in the Coronary Drug Project (Canner 1981, 1983). In this trial, several treatment arms for evaluating cholesterol lowering drugs produced negative trends in the interim results. Through simulation, the probability of achieving a positive or beneficial result was calculated given the observed data at the time of the interim analysis.

Conditional power calculations were utilized in the Vesnarinone in Heart Failure Trial (VEST) (Cohn et al. 1998). In Table 9.2, the test statistics for the logrank test are provided for the information fractions at a series of monitoring committee meetings. Table 9.3 provides conditional power for VEST at three of the interim analyses. A range of intervention effects was used including the beneficial effect (hazard rate less than 1) seen in a previous vesnarinone trial to the observed negative trend (hazard rates of 1.3 and 1.5). It is clear that the conditional power for a beneficial effect was very low by the midpoint of this trial for a null effect or worse. In fact, the conditional power was not encouraging even for the original assumed effect. As described by DeMets et al. (1999) the trial continued beyond this point due to the existence of a previous trial that indicated a large reduction in mortality, rather than the harmful effect observed in VEST.

The Beta-Blocker Heart Attack Trial (DeMets et al. 1984; Beta-Blocker Heart Attack Trial Research Group 1982) also made considerable use of this approach. As discussed, the interim results were impressive with 1 year of follow-up still

Table 9.2 Accumulating results for the Vesnarinone in Heart Failure Trial (Cohn et al. 1998)

Information fraction	Log-rank Z-value (high dose)
0.04	+0.99
0.19	−0.25
0.34	−0.23
0.50	−2.04
0.60	−2.32
0.67	−2.50
0.84	−2.22
0.90	−2.43
0.95	−2.71
1.0	−2.41

Table 9.3 Conditional power for the Vesnarinone in Heart Failure Trial (Cohn et al. 1998)

	Information fraction		
RR	0.50	0.67	0.84
0.50	0.46	<0.01	<0.01
0.70	0.03	<0.01	<0.01
1.0	<0.01	<0.01	<0.01
1.3	<0.01	<0.01	<0.01
1.5	<0.01	<0.01	<0.01

RR relative risk

remaining. One question posed was whether the strong favorable trend ($Z=2.82$) could be lost during that year. The probability of rejecting H_0 at the scheduled end of the trial, given the existing trend (γ_0), was approximately 0.90. This meant that the false positive or Type I error was no more than $\alpha/\gamma_0 = 0.05/0.90$ or 0.056.

9.3 Deciding to Terminate Early

There are five common reasons to terminate a trial earlier than scheduled (Fleming and DeMets 1993; Ellenberg et al. 2003; DeMets et al. 2006; Canner 1981, 1983). First, the trial may show serious adverse effects in the entire intervention group or in a dominating subgroup. Second, the trial may indicate greater than expected beneficial effects. Third, it may become clear that a statistically significant difference by the end of the study is improbable, sometimes referred to as being futile. Fourth, logistical or data quality problem may be so severe that correction is not feasible or participant recruitment is far behind and not likely to achieve the target. Fifth, the question posed may have already been answered elsewhere or may no longer be sufficiently important. A few trials have been terminated because the sponsor decided the trial was no longer a priority but this causes serious ethical dilemmas for investigators and disregards the participants' contribution.

A decision to terminate a study early must be made with caution and in the context of all pertinent data. A number of issues or factors that must be considered thoroughly as part of the decision process include:

1. Possible differences in prognostic factors between the two groups at baseline.
2. Any chance of bias in the assessment of response variables, especially if the trial is not double-blind.
3. The possible impact of missing data. For example, could the conclusions be reversed if the experience of participants with missing data from one group were different from the experience of participants with missing data from the other group?
4. Differential concomitant intervention and levels of participant adherence.
5. Potential adverse events and outcomes of secondary response variables in addition to the outcome of the primary response variable.
6. Internal consistency. Are the results consistent across subgroups and the various primary and secondary outcome measures? In a multicenter trial, the monitoring committee should assess whether the results are consistent across centers. Before stopping, the committee should make certain that the outcome is not due to unusual experience in only one or two centers.
7. In long-term trials, the experience of the study groups over time.
8. The outcomes of similar trials.
9. The impact of early termination on the credibility of the results and acceptability by the clinical community.

The early termination of a clinical trial can be difficult (Fleming and DeMets 1993; Canner 1981, 1983; DeMets 1984, 1990; Freidlin and Korn 2009; Goodman 2009; Montori et al. 2005; Pocock 1992, 2005), not only because the issues involved may be complex and the study complicated but also because the final decision often lies with the consensus of a committee. The statistical methods discussed above are useful guides in this process but should not be viewed as absolute rules. A compilation of diverse monitoring experiences is available (DeMets et al. 2006). A few examples are described here to illustrate key points.

One of the earlier clinical trials conducted in the United States illustrates how controversial the decision for early termination may be. The University Group Diabetes Program (UGDP) was a placebo-control, randomized, double-blind trial designed to test the effectiveness of four interventions used in the treatment of diabetes (Gilbert 1975; Knatterud et al. 1971; Kolata 1979; Meinert et al. 1970). The primary measure of efficacy was the degree of retinal damage. The four interventions were: a fixed dose of insulin, a variable dose of insulin, tolbutamide and phenformin. After the trial was underway, study leaders formed a committee to review accumulating safety data. This committee membership consisted of individuals involved in the UGDP and external consultants. The tolbutamide group was stopped early because the monitoring committee thought the drug could be harmful and did not appear to have any benefit (Meinert et al. 1970). An excess in cardiovascular mortality was observed in the tolbutamide group as compared to the placebo group (12.7% vs. 4.9%) and the total mortality was in the same direction (14.7% vs. 10.2%). Analysis

of the distribution of the baseline factors known to be associated with cardiovascular mortality revealed an imbalance, with participants in the tolbutamide group being at higher risk. This, plus questions about the classification of cause of death, drew considerable criticism. Later, the phenformin group was also stopped because of excess mortality in the control group (15.2% vs. 9.4%) (Gilbert 1975). The controversy led to a further review of the data by an independent group of statisticians. Although they basically concurred with the decisions made by the UGDP monitoring committee (Gilbert 1975), the debate over the study and its conclusion continued (Kolata 1979). This trial certainly highlighted the need for an independent review of the interim data to assess safety.

As discussed earlier, the CDP experience also warns against the dangers of stopping too soon (Canner 1981; The Coronary Drug Project Research Group 1975). In the early months of the study, clofibrate appeared to be beneficial, with the significance level reaching or exceeding 5% on five monitoring occasions (Fig. 9.1). However, because of the repeated testing issue described earlier, the decision was made to continue the study and closely monitor the results. The early difference was not maintained, and at the end of the trial the drug showed no benefit over placebo. It is notable that the mortality curves shown in Fig. 9.2 do not suggest the wide swings observed in the interim analyses shown in Fig. 9.1. The fact that participants were entered over a period of time and thus had various lengths of follow-up at any given interim analysis, explains the difference between the two types of analyses. The decision-making process during the course of the CDP (The Coronary Drug Project Research Group 1970) has been reviewed (DeMets et al. 2006; Canner 1981; The Coronary Drug Project Research Group 1975; The Coronary Drug Project Research Group 1970, 1973). Three of the interventions were terminated early because of potential adverse effects and no apparent benefit. One of the issues in the discontinuation of the high dose estrogen and dextrothyroxine interventions (The Coronary Drug Project Research Group 1970, 1972) concerned subgroups of participants. In some subgroups, the interventions appeared to cause increased mortality, in addition to having a number of other adverse effects. In others, the adverse effects were present, but mortality was only slightly reduced or unchanged. The adverse effects were thought to more than outweigh the minimal benefit in selected subgroups. Also, positive subgroup trends in the dextrothyroxine arm were not maintained over time. After considerable debate, both interventions were discontinued. The low dose estrogen intervention (The Coronary Drug Project Research Group 1973) was discontinued because concerns over major toxicity. Furthermore, it was extremely improbable that a significant difference in a favorable direction for the primary outcome (mortality) could have been obtained had the study continued to its scheduled termination. Using the data available at the time, the number of future deaths in the control group was projected. This indicated that there had to be almost no further deaths in the intervention group for a significance level of 5% to be reached.

Pocock (1992) also warns about the dangers of terminating trials too early for benefit, reflecting on a systematic review of trials stopped early (Montori et al. 2005). At an early interim analysis, the Candesartan in Heart failure Assessment of Reduction in Mortality and Morbidity (CHARM) trial (Pocock et al. 2005) had a 25% mortality

benefit ($p<0.001$) from candesartan compared to a placebo control, but for a variety of reasons the trial continued and found after a median of 3 years of follow-up only a 9% nonsignificant difference in mortality. Continuing the trial revealed that the early mortality benefit was probably exaggerated and allowed other long-term intervention effects to be discovered. In general, trials stopped early for benefit often do not report in sufficient detail the rationale for early termination and often show implausibly large intervention effects based on only a small number of events (Freidlin and Korn 2009). This phenomenon is well recognized (Goodman 2009). Thus, while there are sound ethical reasons to terminate trials early because of benefit, these decisions must be cautioned by many examples showing that early trends not being reliable or sustainable. Nevertheless, there is a natural tension between getting the estimate of treatment benefit precise and allowing too many participants to be exposed to the inferior intervention (Freidlin and Korn 2009). Statistical methods are useful as guidelines but not adequate as rules and the best approach based on experience is to utilize a properly constituted monitoring committee, charged with weighing the benefits and risks of early termination.

Some of the most challenging monitoring scenarios involve an emerging negative trend for the primary outcomes. The PROMISE and PROFILE experiences described below illustrate this, but they are not unique (DeMets et al. 1999; Furberg et al. 1993; Pater 1994; Swedberg et al. 1992; Sylvester et al. 1994). Trials with persistent nonsignificant negative trends may have no real chance of reversing and indicating a benefit from intervention. In some circumstances, that observation may be sufficient to end the trial since if a result falls short of establishing benefit, the intervention would not be used. For example a new expensive or invasive intervention would likely need to be more effective than a standard intervention to be used. In other circumstances, a neutral result may be important, so a small negative trend, still consistent with a neutral result, would argue for continuation. If a treatment is already in clinical use on the basis of other indications, as in the case of the drugs used in PROMISE and PROFILE, an emerging negative trend may not be sufficient evidence to alter clinical practice. If a trial terminates early without resolving convincingly the harmful effects of an intervention, that intervention may still continue to be used. This practice would put future patients at risk, and perhaps even participants in the trial as they return to their usual healthcare system. In that case, the investment of participants, investigators, and sponsors would not have resolved an important question. There is a serious and delicate balance between the responsibility to safeguard the participants in the trial and the responsibility for all concurrent and future patients (DeMets et al. 1999).

The Cardiac Arrhythmia Suppression Trial (CAST) was a multicenter randomized double blind placebo-controlled trial evaluating the effects of three drugs on total mortality and sudden death (The Cardiac Arrhythmia Suppression Trial (CAST) Investigators 1989). Epidemiological data showed an association between the presence of irregular ventricular heartbeats or arrhythmias and the incidence of sudden death, presumably due to serious arrhythmias. Encainide, flecainide, moricizine were among the drugs developed to suppress such arrhythmias and they became widely used after approval by the drug regulatory agency for that indication. At the first

monitoring committee review, a mortality trend began to appear but the number of events was relatively small (Friedman et al. 1993). Because the monitoring committee decided no definitive conclusion could be reached on the basis of so few events, it elected to remain blinded to the treatment assignment. However, before the next scheduled meeting, the statistical center alerted the committee that the trends continued and were now nearing the CAST monitoring criteria for stopping. In a conference call meeting, the monitoring committee became unblinded and learned that the trends were in the unexpected direction, that is, toward harm from the active treatment. A number of confirmatory and exploratory analyses were requested by the monitoring committee and a meeting was held a few weeks later to discuss fully these unexpected results. After a thorough review, results were consistent across outcome variables and participant subgroups, no biases could be identified which would explain these result, and the encainide and flecainide arms of the trial were terminated after only 15% of the expected mortality events observed because of an adverse effect (63 deaths in the two active arms vs. 26 deaths in the corresponding placebo arms). The third arm (moricizine) continued since there were no convincing trends at that time, but it too was eventually stopped due to adverse experiences (The Cardiac Arrhythmia Suppression Trial II Investigators 1992). The CAST experience points out that monitoring committees must be prepared for the unexpected and that large trends may emerge quickly. Even in this dramatic result, the decision was not simple or straightforward. Many of the issues discussed earlier were covered thoroughly before a decision was reached (Friedman et al. 1993).

The Women's Health Initiative (WHI) was one of the largest and most complex trials ever conducted, certainly in women (The Women's Health Initiative Steering Committee 2004; Writing Group for the Women's Health Initiative Investigators 2002). This partial factorial trial evaluated three interventions in postmenopausal women: (1) hormone replacement therapy (HRT), (2) a low fat diet, and (3) calcium and vitamin D supplementation. Each intervention, in principle, could affect multiple organ systems, each with multiple outcomes. For example, HRT was being evaluated for its effect on cardiovascular events such as mortality and fatal and nonfatal myocardial infarction. HRT can also affect bone density, the risk of fracture, and breast cancer. The HRT component was also stratified into those with an intact uterus, who received both estrogen and progestin, and those without a uterus who received estrogen alone. The estrogen–progestin arm was terminated early due to increases in deep vein thrombosis, pulmonary embolism, stroke, and breast cancer and a trend toward increased heart disease as shown in Fig. 9.8 although there was a benefit in bone fracture as expected (Writing Group for the Women's Health Initiative Investigators 2002). There was no observed difference in total mortality or the overall global index, the composite outcome defined in the protocol, as shown in Fig. 9.9. The WHI is an excellent example of the challenges of monitoring trials with composite outcomes where component trends are not consistent. In such cases, the most important or most clinically relevant component may have to dominate in the decision process, even if not completely specified in the protocol or the monitoring committee charter. Later, the WHI estrogen-alone arm was also terminated, primarily due to increased pulmonary embolus and stroke, though there was no difference

in myocardial infarction or total mortality (The Women's Health Initiative Steering Committee 2004). The formal monitoring process had to account for multiple interventions, multiple outcomes and repeated testing.

The Justification for the Use of Statin in Prevention: An Intervention Trial Evaluating Rosuvastatin (JUPITER) trial compared a statin agent, which lowers both LDL cholesterol and C-reactive protein, in 17,802 patients with elevated high-sensitivity C-reactive protein levels but without hyperlipidemia (Ridker et al. 2008). The primary outcome was the occurrence of the combination of myocardial infarction, stroke, arterial revascularization, hospitalization for unstable angina, or death from cardiovascular causes. The trial, which was stopped about 3 years early after 2 years of follow-up, found a clear lowering of both LDL and C-reactive protein and demonstrated a corresponding reduction in the primary outcome (hazard ratio (HR) of 0.56, $p < 0.00001$). Similar reductions were observed for myocardial infarction (HR, 0.46), for stroke (HR, 0.52), for revascularization or unstable angina (HR, 0.53), for the combined end point of myocardial infarction, stroke, or death from cardiovascular causes (HR, 0.53), and for death from any cause (HR, 0.80), all being statistically significant. In addition, all of the major predefined subgroups were consistent. Still, there was criticism that the cardiovascular mortality was not significant even though overall mortality was (Ridker 2009; Voss et al. 2009). This raises the difficult question when using combined outcomes as the primary if each component or at least some components should also be statistically significant before terminating a trial. In general, trials are not designed to demonstrate statistically significant results for any of the components usually due to low events for each of them. To do so would require trials much larger than the one designed. If a component of the combined outcome is of paramount importance, then that outcome should be established as the primary and the trial designed accordingly. In the case of the JUPITER trial, the results for the primary outcome and nearly all of its components as well as overall mortality appear to be compelling for a trial to be terminated early. This is especially the case when total mortality is significantly reduced in addition to the primary. Another approach to a focus on a component of the primary outcome was in the CHARM program, in which three trials that comprised the overall program each had cardiovascular death and heart failure hospitalization as its primary outcome, and the overall program was powered to assess all-cause mortality. The monitoring committee focused on the effect on mortality in the overall program as the criterion for early termination (Ridker 2009).

In some instances, a trial may be terminated because the hypothesis being tested has been convincingly answered by other ongoing trials. This was the case with trials evaluating warfarin in the treatment of atrial fibrillation (Tegeler and Furberg 2006). Between 1985 and 1987, five trials were launched to evaluate warfarin to prevent strokes in participants with atrial fibrillation. Three of the trials were terminated early by 1990, reporting significant reductions in embolic complications. One of the remaining trials was also terminated early, largely due to the ethical aspects of continuing trials when the clinical question being tested has already been answered. The window of opportunity to further evaluate the intervention had closed.

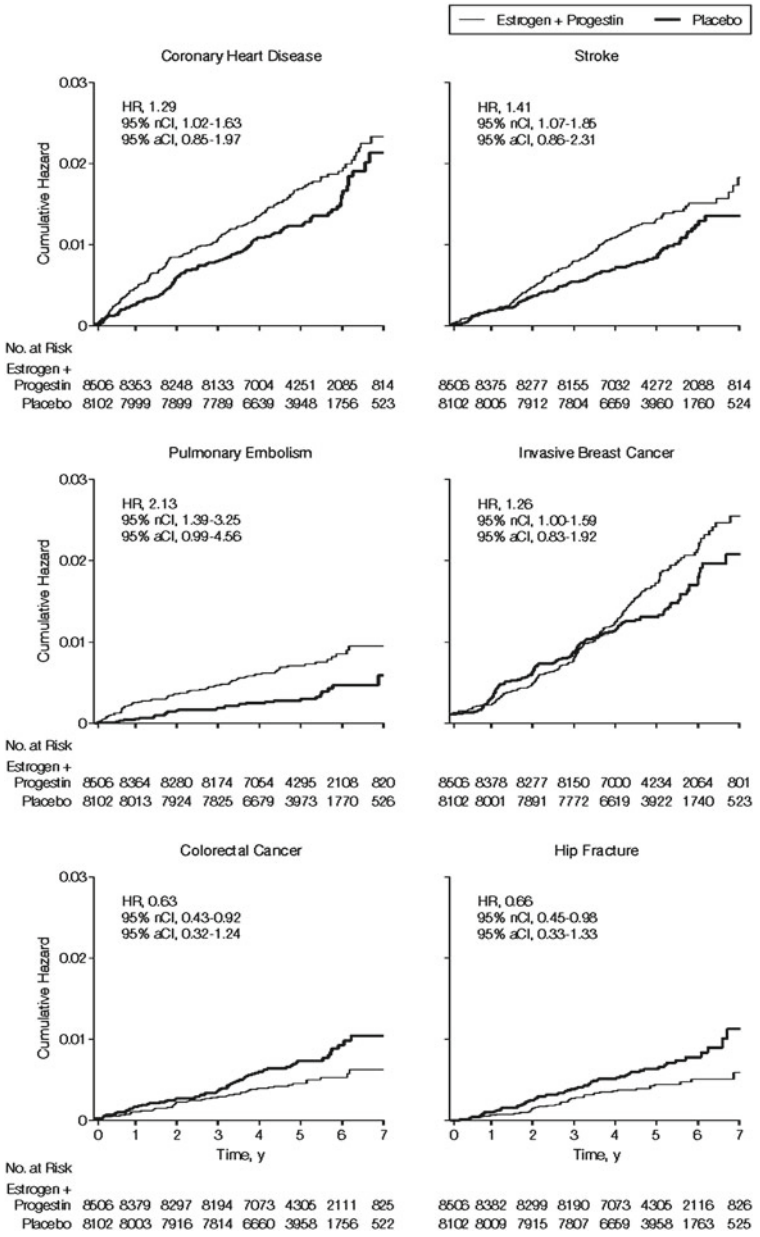


Fig. 9.8 WHI Kaplan-Meier Estimates of Cumulative Hazards for Selected Clinical Outcomes (Writing Group for the Women's Health Initiative Investigators 2002). HR = hazard ratio; nCI = nominal confidence interval; aCI = adjusted confidence interval

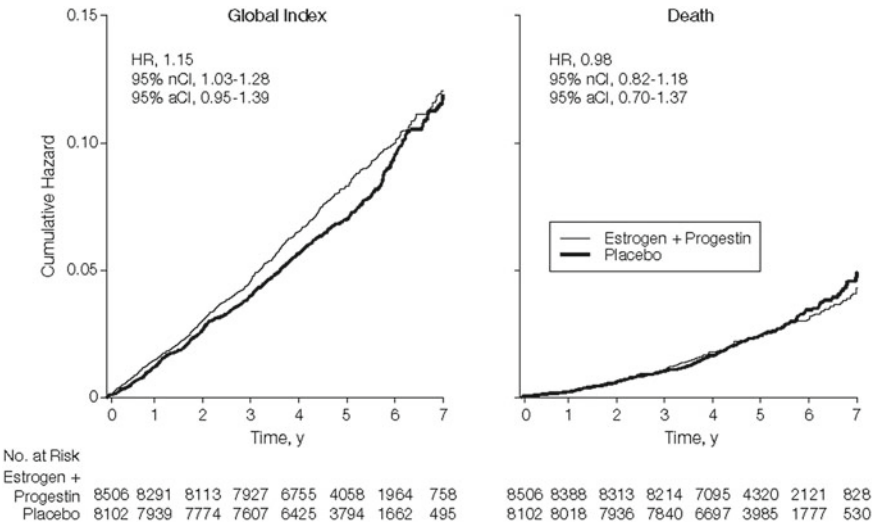


Fig. 9.9 WHI Kaplan-Meier Estimates of Cumulative Hazards for Global Index and Death (Writing Group for the Women's Health Initiative Investigators 2002). HR = hazard ratio; nCI = nominal confidence interval; aCI = adjusted confidence interval

As we have already discussed, the decision to terminate a trial is complex. It is never based on a single outcome and may require more than one monitoring committee meeting before a recommendation to terminate is reached. Timing of the recommendation can also be questioned by those external to the trial. In the Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events (ILLUMINATE) trial (Barter et al. 2007), a new agent torcetrapib, a cholesterylester transfer protein inhibitor that increases HDL cholesterol, was tested to reduce major cardiovascular events. ILLUMINATE was a randomized, double-blind study involving 15,067 patients at high cardiovascular risk, receiving either torcetrapib plus atorvastatin (a statin which lowers LDL cholesterol) or atorvastatin alone. The primary outcome was defined as time to death from coronary heart disease, nonfatal myocardial infarction, stroke, or hospitalization for unstable angina, whichever occurred first. ILLUMINATE clearly demonstrated an increase in HDL, which would be expected to cause a reduction in cardiovascular risk. However, the trial was terminated early by the monitoring committee and the investigators because of an increased risk of death and cardiac events in patients receiving torcetrapib (Barter et al. 2007). To conclude that torcetrapib improved HDL but caused harmful clinical effects was of course disappointing since this was the first testing of an exciting new class of drugs. However, the timing of the recommendation to terminate was challenged by a regulatory agency, which recognized the complexity of such decisions but argued that the trial could and perhaps should have been terminated earlier (Hedenmalm et al. 2008). Determining at what point there is sufficient and compelling evidence to make a recommendation for termination is often challeng-

ing. Monitoring committees do not have the benefit of hindsight while in process of monitoring a trial.

On occasion, trials may have achieved a significant benefit, or show strong trends for benefit, but the monitoring committee recommended early termination for safety reasons. Two trials, the Thrombin Receptor Antagonist in Secondary Prevention of Atherothrombotic Ischemic Events (TRA 2P) trial (Morrow et al. 2012) and the Thrombin Receptor Antagonist for Clinical Event Reduction in Acute Coronary Syndrome (TRACER) trial (Tricoci et al. 2011) provide examples of such instances. Both trials evaluated a new platelet inhibition agent vorapaxar compared with placebo. TRA 2P had the primary outcome as a composite of death from cardiovascular causes, myocardial infarction, or stroke. TRACER had a composite outcome of death from cardiovascular causes, myocardial infarction, stroke, recurrent ischemia with rehospitalization, or urgent coronary revascularization. Both trials, TRA 2P with 26,449 patients and TRACER with 12,944 patients, had statistically significant beneficial effects in their respective primary outcomes (HR of 0.87 and 0.89). In TRACER, there were 1031 primary events in the treated patients and 1102 in the placebo controls. The secondary composite of cardiovascular death, MI and stroke had 822 vs 910 events ($p = 0.02$). However, the rates of intracranial bleeding was 1.2% versus 0.2% yielding a hazard ratio of 3.39 ($p < 0.001$). The monitoring committees for both trials decided that the serious bleeding risks overwhelmed any emerging benefits and recommended early termination and/or modification of the protocol for unacceptable bleeding complications including intracranial hemorrhage.

In all of these studies, the decisions were difficult and involved many analyses, thorough review of the literature, and an understanding of the biological processes. As described above, a number of questions must be answered before serious consideration should be given to early termination. As noted elsewhere, the relationship between clinical trials and practice is very complex and this complexity is evident in the monitoring process (Liberati 1994; O'Neill 1994).

References

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association, 58*(302), 365–383.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, 9–26.

Armitage, P. (1975). *Sequential medical trials* (2nd ed.). New York, NY: Wiley.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society Series A (General), 132*(2), 235–244.

Barter, P. J., Caulfield, M., Eriksson, M., Grundy, S. M., Kastelein, J. J. P., Komajda, M., et al. (2007). Effects of torcetrapib in patients at high risk for coronary events. *New England Journal of Medicine, 357*(21), 2109–2122.

Baum, M., Houghton, J., & Abrams, K. (1994). Early stopping rules-clinical perspectives and ethical considerations. *Statistics in Medicine, 13*(13), 1459–1469.

Beta-Blocker Heart Attack Trial Research Group. (1982). A randomized trial of propranolol in patients with acute myocardial infarction: I. Mortality results. *JAMA, 247*(12), 1707–1714.

Bross, I. (1952). Sequential medical plans. *Biometrics, 8*(3), 188–205.

Canner, P. L. (1981). Practical aspects of decision-making in clinical trials-the Coronary Drug Project as a case-study. *Controlled Clinical Trials, 1*(4), 363–376.

Canner, P. L. (1983). Monitoring of the data for evidence of adverse or beneficial treatment effects. *Controlled Clinical Trials, 4*(4), 467–483.

Chang, M. N., & O'Brien, P. C. (1986). Confidence intervals following group sequential tests. *Controlled Clinical Trials, 7*(1), 18–26.

Cohn, J. N., Goldstein, S. O., Greenberg, B. H., Lorell, B. H., Bourge, R. C., Jaski, B. E., et al. (1998). A dose-dependent increase in mortality with vesnarinone among patients with severe heart failure. *New England Journal of Medicine, 339*(25), 1810–1816.

Crowley, J., Green, S., Liu, P. Y., & Wolf, M. (1994). Data monitoring committees and early stopping guidelines: The Southwest Oncology Group experience. *Statistics in Medicine, 13*(13–14), 1391–1399.

DeMets, D. L. (1984). Stopping guidelines versus stopping rules—A practitioners point of view. *Communications in Statistics-Theory and Methods, 13*(19), 2395–2417.

DeMets, D. L. (1987). Practical aspects in data monitoring: A brief review. *Statistics in Medicine, 6*(7), 753–760.

DeMets, D. L. (1990). Data monitoring and sequential analysis-An academic perspective. *JAIDS Journal of Acquired Immune Deficiency Syndromes, 3,* S124–S133.

DeMets, D. L. (2006). Futility approaches to interim monitoring by data monitoring committees. *Clinical Trials, 3*(6), 522–529.

DeMets, D. L., Fleming, T. R., Whitley, R. J., Childress, J. F., Ellenberg, S. S., Foulkes, M., et al. (1995). The data and safety monitoring board and acquired immune deficiency syndrome (AIDS) clinical trials. *Controlled Clinical Trials, 16*(6), 408–421.

DeMets, D. L., Furberg, C., & Friedman, L. M. (2006). *Data monitoring in clinical trials: A case studies approach* (p. 2006). New York, NY: Springer.

DeMets, D. L., Hardy, R., Friedman, L. M., & Gordon Lan, K. K. (1984). Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Controlled Clinical Trials, 5*(4), 362–372.

DeMets, D. L., & Lan, K. K. G. (1989). Discussion of: Interim analyses: The repeated confidence interval approach by C. Jennison and BW Turnbull. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 51,* 344.

DeMets, D. L., & Lan, K. K. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine, 13*(13–14), 1341–1352.

DeMets, D. L., & Lan, G. (1995). *The alpha spending function approach to interim data analyses. Recent advances in clinical trial design and analysis* (pp. 1–27). Berlin: Springer.

DeMets, D. L., Pocock, S. J., & Julian, D. G. (1999). The agonising negative trend in monitoring of clinical trials. *The Lancet., 354*(9194), 1983–1988.

DeMets, D. L., & Ware, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika, 67*(3), 651–660.

DeMets, D. L., & Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical-trials. *Biometrika, 69*(3), 661–663.

Ellenberg, S. S., Fleming, T. R., & DeMets, D. L. (2003). *Data monitoring committees in clinical trials: A practical perspective*. New York: Wiley.

Emerson, S. S., & Fleming, T. R. (1989). Symmetric group sequential test designs. *Biometrics, 45*(3), 905–923.

Emerson, S. S., & Fleming, T. R. (1990a). Interim analyses in clinical trials. *Oncology (Williston Park, NY), 4*(3), 126.

Emerson, S. S., & Fleming, T. R. (1990b). Parameter estimation following group sequential hypothesis testing. *Biometrika, 77*(4), 875–892.

Falissard, B., & Lellouch, J. (1992). A new procedure for group sequential analysis in clinical trials. *Biometrics*, 373–388.

Feyzi, J., Julian, D. G., Wikstrand, J., & Wedel, H. (2006). *Data monitoring experience in the Metoprolol CR/XL randomized intervention trial in chronic heart failure: Potentially high-risk treatment in high-risk patients* (pp. 136–147). Data Monitoring in Clinical Trials: Springer.

Fisher, M. R., Roecker, E. B., & DeMets, D. L. (2001). The role of an independent statistical analysis center in the industry-modified National Institutes of Health model. *Drug Information Journal, 35*(1), 115–129.

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

Fleming, T. R., & Watelet, L. F. (1989). Approaches to monitoring clinical trials. *Journal of the National Cancer Institute, 81*(3), 188–193.

Freedman, B. (1987). Equipoise and the ethics of clinical research. *The New England Journal of Medicine*.

Freedman, L. S., Lowe, D., & Macaskill, P. (1983). Stopping rules for clinical trials. *Statistics in Medicine, 2*(2), 167–174.

Freidlin, B., & Korn, E. L. (2009). Stopping clinical trials early for benefit: Impact on estimation. *Clinical Trials, 6*(2), 119–125.

Friedman, L. M., Bristow, J. D., Hallstrom, A., Schron, E., Proschan, M., & Verter, J. et al. (1993). Data monitoring in the cardiac arrhythmia suppression trial. *Online Journal of Current Clinical Trials*, 79.

Furberg, C. D., Campbell, R., & Pitt, B. (1993). ACE inhibitors after myocardial infarction. *New England Journal of Medicine, 328*(13), 966–969.

Gange, S. J., & DeMets, D. L. (1996). Sequential monitoring of clinical trials with correlated responses. *Biometrika, 83*(1), 157–167.

Geller, N. L. (1994). Discussion of "Interim analysis: The alpha spending approach". *Statistics in Medicine, 13*(13–14), 1353–1356.

Gilbert, J. P. (1975). Report of the committee for the assessment of biometric aspects of controlled trials of hypoglycemic agents. *JAMA, 231*(6), 583–608.

Goodman, S. N. (2009). Stopping trials for efficacy: An almost unbiased view. *Clinical Trials, 6*(2), 133–135.

Gould, A. L., & Pecore, V. J. (1982). Group sequential methods for clinical trials allowing early acceptance of Ho and incorporating costs. *Biometrika, 69*(1), 75–80.

Green, S., & Crowley, J. (1993). Data monitoring committees for Southwest Oncology Group clinical trials. *Statistics in Medicine, 12*(5–6), 451–455.

Halperin, M., Gordon Lan, K. K., Ware, J. H., Johnson, N. J., & DeMets, D. L. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials, 3*(4), 311–323.

Harrington, D., Crowley, J., George, S. L., Pajak, T., Redmond, C., & Wieand, S. (1994). The case against independent monitoring committees. *Statistics in Medicine, 13*(13–14), 1411–1414.

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *The British Journal of Radiology, 44*(526), 793–797.

Heart Special Project Committee. (1988). Organization, review, and administration of cooperative studies (Greenberg Report): A report from the Heart Special Project Committee to the National Advisory Heart Council, May 1967. *Controlled Clinical Trials, 9*(2), 137–148.

Hedenmalm, K., Melander, H., & Alvan, G. (2008). The conscientious judgement of a DSMB—Statistical stopping rules re-examined. *European Journal of Clinical Pharmacology, 64*(1), 69–72.

Hughes, M. D., & Pocock, S. J. (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine, 7*(12), 1231–1242.

Hwang, I. K., Shih, W. J., & De Cani, J. S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine, 9*(12), 1439–1445.

Jennison, C., & Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science*, 299–317.

Jennison, C., & Turnbull, B. W. (1984). Repeated confidence-intervals for group sequential clinical-trials. *Controlled Clinical Trials, 5*(1), 33–45.

Jennison, C., & Turnbull, B. W. (1989). Interim analyses: The repeated confidence interval approach. *Journal of the Royal Statistical Society: Series B (Methodological), 51*(3), 305–361.

Jennison, C., & Turnbull, B. W. (1997). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association, 92*(440), 1330–1341.

Jennison, C., & Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. London: Taylor & Francis.

Kim, K. (1989). point estimation following group sequential tests. *Biometrics, 45*(2), 613–617.

Kim, K., & DeMets, D. L. (1987a). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika, 74*(1), 149–154.

Kim, K., & DeMets, D. L. (1987b). Confidence intervals following group sequential tests in clinical trials. *Biometrics, 43*(4), 857–864.

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

Knatterud, G. L., Meinert, C. L., Klimt, C. R., Osborne, R. K., & Martin, D. B. (1971). Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: IV. A preliminary report on phenformin results. *JAMA, 217*(6), 777–784.

Kolata, G. B. (1979). Controversy over study of diabetes drugs continues for nearly a decade. *Science (New York, NY), 203*(4384), 986.

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical-trials. *Biometrika, 70*(3), 659–663.

Lan, K. K. G., & DeMets, D. (1989a). Group sequential procedures: Calendar versus information time. *Statistics in Medicine, 8*(10), 1191–1198.

Lan, K. K. G., & DeMets, D. L. (1989b). Changing frequency of interim analysis in sequential monitoring. *Biometrics, 45*(3), 1017–1020.

Lan, K. K. G., DeMets, D. L., & Halperin, M. (1984). More flexible sequential and non-sequential designs in long-term clinical-trials. *Communications in Statistics-Theory and Methods, 13*(19), 2339–2353.

Lan, K. K. G., & Lachin, J. M. (1990). Implementation of group sequential logrank tests in a maximum duration trial. *Biometrics, 46*(3), 759–770.

Lan, K. K. G., Reboussin, D. M., & DeMets, D. L. (1994). Information and information fractions for design and sequential monitoring of clinical-trials. *Communications in Statistics-Theory and Methods, 23*(2), 403–420.

Lan, K. K. G., Rosenberger, W. F., & Lachin, J. M. (1993). Use of spending functions for occasional or continuous monitoring of data in clinical trials. *Statistics in Medicine, 12*(23), 2219–2231.

Lan, K. K. G., Simon, R., & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis, 1*(3), 207–219.

Lan, K. K. G., & Wittes, J. (1988). The B-value: A tool for monitoring data. *Biometrics, 44*(2), 579–585.

Lan, K. K. G., & Zucker, D. M. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion. *Statistics in Medicine, 12*(8), 753–765.

Lee, J. W. (1994). Group sequential testing in clinical trials with multivariate observations: A review. *Statistics in Medicine, 13*(2), 101–111.

Lee, J. W., & DeMets, D. L. (1991). Sequential comparison of changes with repeated measurements data. *Journal of the American Statistical Association, 86*(415), 757–762.

Lee, J. W., & DeMets, D. L. (1992). Sequential rank-tests with repeated measurements in clinical-trials. *Journal of the American Statistical Association., 87*(417), 136–142.

Li, Z. Q., & Geller, N. L. (1991). On the choice of times for data analysis in group sequential clinical trials. *Biometrics, 47*(2), 745–750.

Liberati, A. (1994). Conclusions. 1: The relationship between clinical trials and clinical practice: The risks of underestimating its complexity. *Statistics in Medicine, 13*(13–14), 1485–1491.

McPherson, C. K., & Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society Series A (General)*, 15–25.

Meinert, C. L. (1998). Masked monitoring in clinical trials-blind stupidity? *New England Journal of Medicine, 338*(19), 1381–1382.

Meinert, C. L., Knatterud, G. L., Prout, T. E., & Klimt, C. R. (1970). A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. II. *Mortality results Diabetes, 19*(Suppl 2), 789–830.

Montori, V. M., Devereaux, P. J., Adhikari, N. K., Burns, K. E., Eggert, C. H., Briel, M., et al. (2005). Randomized trials stopped early for benefit: A systematic review. *Journal of the American Medical Association, 294*(17), 2203.

Morrow, D. A., Braunwald, E., Bonaca, M. P., Ameriso, S. F., Dalby, A. J., Fish, M. P., et al. (2012). Vorapaxar in the secondary prevention of atherothrombotic events. *New England Journal of Medicine, 366*(15), 1404–1413.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35*(3), 549–556.

O'Neill, R. T. (1994). Conclusions. 2: The relationship between clinical trials and clinical practice: The risks of underestimating its complexity. *Statistics in Medicine, 13*(13–14), 1493–1499.

Pater, J. L. (1994). Timing the collaborative analysis of three trials comparing 5-U plus folinic acid (FUFA) to surgery alone in the management of resected colorectal cancer: A national cancer institute of canada clinical trials group (NCIC-CTG) perspective. *Statistics in Medicine, 13*(13–14), 1337–1340.

Pawitan, Y., & Hallstrom, A. (1990). Statistical interim monitoring of the cardiac arrhythmia suppression trial. *Statistics in Medicine, 9*(9), 1081–1090.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., et al. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer, 34*(6), 585.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*(2), 191–199.

Pocock, S. J. (1978). Size of cancer clinical trials and stopping rules. *British Journal of Cancer, 38*(6), 757.

Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics, 38*(1), 153–162.

Pocock, S. J. (1992). When to stop a clinical trial. *BMJ. British Medical Journal, 305*(6847), 235.

Pocock, S. J. (2005). When (not) to stop a clinical trial for benefit. *JAMA, 294*(17), 2228–2230.

Pocock, S. J., & Hughes, M. D. (1989). Practical problems in interim analyses, with particular regard to estimation. *Controlled Clinical Trials, 10*(4 Suppl), 209S–221S.

Pocock, S. J., Wang, D., Wilhelmsen, L., & Hennekens, C. H. (2005). The data monitoring experience in the Candesartan in Heart Failure Assessment of Reduction in Mortality and morbidity (CHARM) program. *American Heart Journal, 149*(5), 939–943.

Proschan, M. A., Follmann, D. A., & Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 1131–1143.

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. New York: Springer.

Reboussin, D. M., DeMets, D. L., Kim, K., & Lan, K. K. G. Lan-DeMets Method - Statistical Programs for Clinical Trials 2003 [updated 17 November 2003. 2.1: Available from: https://www.biostat.wisc.edu/content/lan-demets-method-statistical-programs-clinical-trials.

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

Ridker, P. M. (2009). The JUPITER trial: Results, controversies, and implications for prevention. *Circulation: Cardiovascular Quality and Outcomes, 2*(3), 279–285.

Ridker, P. M., Danielson, E., Fonseca, F. A., Genest, J., Gotto, A. M., Jr., Kastelein, J. J., et al. (2008). Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *New England Journal of Medicine, 359*(21), 2195.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society, 58*(5), 527–535.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 1397–1409.

Rosner, G. L., & Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika, 75*(4), 723–729.

Scharfstein, D. O., Tsiatis, A. A., & Robins, J. M. (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association, 92*(440), 1342–1350.

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika, 65*(2), 341–349.

Su, J. Q., & Lachin, J. M. (1992). Group sequential distribution-free methods for the analysis of multivariate observations. *Biometrics, 48*(4), 1033–1042.

Swedberg, K., Held, P., Kjekshus, J., Rasmussen, K., Ryden, L., & Wedel, H. (1992). Effects of the early administration of enalapril on mortality in patients with acute myocardial infarction: results of the Cooperative New Scandinavian Enalapril Survival Study II (CONSENSUS II). *New England Journal of Medicine, 327*(10), 678–684.

Sylvester, R., Bartelink, H., & Rubens, R. (1994). A reversal of fortune: Practical problems in the monitoring and interpretation of an EORTC breast cancer trial. *Statistics in Medicine, 13*(13–14), 1329–1335.

Tegeler, C. H., & Furberg, C. D. (2006). *Lessons from warfarin trials in atrial fibrillation: Missing the window of opportunity. Data monitoring in clinical trials* (pp. 312–319). Berlin: Springer.

The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. (1989). Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *New England Journal of Medicine, 321*(6), 406–412.

The Coronary Drug Project Research Group. (1970). The Coronary Drug Project: Initial findings leading to modifications of its research protocol. *JAMA, 214*(7), 1303–1313.

The Coronary Drug Project Research Group. (1972). The Coronary Drug Project: Findings leading to further modifications of its protocol with respect to dextrothyroxine. *JAMA, 220*(7), 996–1008.

The Coronary Drug Project Research Group. (1973). The Coronary Drug Project: Findings leading to discontinuation of the 2.5-mg/day estrogen group. *JAMA, 226*(6), 652–657.

The Coronary Drug Project Research Group. (1975). Clofibrate and niacin in coronary heart disease. *JAMA, 231*(4), 360–381.

The Cardiac Arrhythmia Suppression Trial II Investigators. (1992). Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. *New England Journal of Medicine, 327*(4), 227–233.

The Women's Health Initiative Steering Committee. (2004). Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: The Women's Health Initiative randomized controlled trial. *JAMA, 291*(14), 1701–1712.

Tricoci, P., Huang, Z., Held, C., Moliterno, D. J., Armstrong, P. W., Van de Werf, F., et al. (2011). Thrombin-receptor antagonist vorapaxar in acute coronary syndromes. *New England Journal of Medicine, 366*(1), 20–33.

Tsiatis, A. A., Rosner, G. L., & Mehta, C.R. (1984). Exact confidence intervals following a group sequential test. *Biometrics*, 797–803.

Voss, E., Rose, C. P., & Biron, P. (2009). JUPITER, a statin trial without cardiovascular mortality benefit. *Circulation: Cardiovascular Quality and Outcomes, 2*(3), 279–285.

Wald, A. (2013). *Sequential analysis*. New York, NY: Dover.

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

Wei, L. J., Su, J. Q., & Lachin, J. M. (1990). Interim analyses with repeated measurements in a sequential clinical-trial. *Biometrika, 77*(2), 359–364.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika, 73*(3), 573–581.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. New York, NY: Wiley.

Whitehead, J. (1999). On being the statistician on a Data and Safety Monitoring Board. *Statistics in Medicine, 18*(24), 3425–3434.

Whitehead, J., & Facey, K. M. (1991). *Analysis after a sequential trial: A comparison of orderings of the sample space*. Brussels: Joint Society for Clinical Trials/International Society for Clinical Biostatistics.

Writing Group for the Women's Health Initiative Investigators. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal results from the Women's Health Initiative randomized controlled trial. *JAMA, 288*(3), 321–333.

Wu, M. C., & Gordon Lan, K. K. (1992). Sequential monitoring for comparison of changes in a response variable in clinical studies. *Biometrics, 48*(3), 765–779.

Chapter 10
Design and Data Analysis of Multiregional Clinical Trials (MRCTs)—Theory and Practice

Chi-Tian Chen, Hsiao-Hui Tsou, Jung-Tzu Liu, Chin-Fu Hsiao, Fei Chen, Gang Li and K. K. G. Lan

10.1 Introduction

In recent years, multiregional clinical trial (MRCT) has become a preferred strategy to develop new medicines. Implementing the same protocol to include subjects from many geographical regions around the world, MRCTs could speed up the patient enrollment, thus resulted in a quicker drug development and obtain faster approval of the drug globally. At the same time, the MRCT strategy is expected to maintain the sample size at the similar level, i.e., without significantly driving up the cost and slowing down the speed of the development. As the draft ICH E17 (2016) states: 'The underlying assumption of the conduct of MRCTs is that the treatment effect is clinically meaningful and relevant to all regions being studied', which is often referred to as consistency among regions. The proper consistency assessment is closely related the sample size and the number of regions as well as the approach

Chi-Tian Chen, Hsiao-Hui Tsou: *These authors contributed equally to this research.*

C.-T. Chen · H.-H. Tsou · C.-F. Hsiao
Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Miaoli County, Taiwan, Republic of China

H.-H. Tsou
Graduate Institute of Biostatistics, College of Public Health, China Medical University, Taichung, Taiwan, Republic of China

J.-T. Liu
Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan, Republic of China

F. Chen · G. Li (✉) · K. K. G. Lan (✉)
Janssen R & D, Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA
e-mail: GLi@its.jnj.com

K. K. G. Lan
e-mail: Glan@its.jnj.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_10

to combine the treatment effects among regions. Thus, how to combine evidence of treatment effect from different regions is an increasingly important topic for consideration in clinical drug development. In addition, it is of interest to explore how to incorporate the possible regional differences into trial planning. Three models are proposed in literature for combining treatment effects from different regions: the fixed effect model (FEM); the continuous random effects model (CREM); and the discrete random effects model (DREM).

In this chapter we will discuss these three models, focusing on comparison of DREM and CREM and their operational characteristics in MRCTs. And review the drop-min approach for analyzing the data when a region in an MRCT is considered as inconsistent to the others.

10.2 FEM, CREM and DREM Models

Let us first introduce notations for the three models. Suppose an MRCT is conducted in M regions for comparing a test product, T, with a placebo control, C. The sample sizes for groups T and C at region i are N_{Ti} and N_{Ci}, respectively. The total sample sizes for groups T and C are $N_T = \sum_{i=1}^{M} N_{Ti}$ and $N_C = \sum_{i=1}^{M} N_{Ci}$, respectively. We assume that $N_T = N_C = N$, and that $N_{Ti} = N_{Ci} = N_i$. Let X_{Tij} and X_{Cij} be the responses the for the jth subject in ith region of T and C, respectively. They can be expressed as $X_{Tij} = \mu_{Tij} + \varepsilon_{Tij}$, and $X_{Cij} = \mu_{Cij} + \varepsilon_{Cij}$ where ε_{Tij} and ε_{Cij} are assumed to be independent normally distributed with mean 0 and variance σ^2, $i = 1, 2, \ldots, M$, $j = 1, \ldots, N_i$. We denote the treatment effect in the ith region as $v_i = \mu_{Ti} - \mu_{Ci}$, for $i = 1, 2, \ldots, M$. Naturally, v_i is estimated by $\hat{v}_i = \overline{X}_{Ti} - \overline{X}_{Ci}$, where $\overline{X}_{Ti}$ and $\overline{X}_{Ci}$ are the sample means of T and C, respectively. These three models treat v_i in three different ways as we will illustrate in next subsections.

10.2.1 Fixed Effect Model

Traditionally, a common treatment effect and an equal variability of the primary endpoint across regions are assumed for the design and evaluation of MRCTs, such as an approach to rationalize partitioning the total sample size among the regions (Kawai et al. 2008), consistency criteria approach (Ko et al. 2010), statistical consideration from an Asian perspective (Tsou et al. 2010), similarity assessment using Bayesian most plausible prediction (Tsou et al. 2011), and a consistency approach across all participating regions (Tsou et al. 2012).

A fixed effect model (FEM) follows the traditional approach and assumes that all regional treatment effects are equal, i.e., $v_1 = \cdots = v_M = v$. And v is called overall treatment effect. Let $\hat{v}_i$ be the estimate of v_i, We are interested in testing the following hypothesis of overall treatment effect

$$H_0 : v = 0 \text{ versus } H_A : v > 0. \tag{10.1}$$

Then we have a standard two-sample Z test

$$Z_{\text{FEM}} = \frac{\hat{v}_{\text{FEM}}}{\sigma\sqrt{1/\text{N} + 1/\text{N}}} \tag{10.2}$$

where $\hat{v}_{FEM} = \sum_{i=1}^{M} N_i \hat{v}_i / N$ is the estimate of v under FEM. The null hypothesis would be rejected at the significance level α if the test statistic $Z_{\text{FEM}} > z_{1-\alpha}$, where $z_{1-\alpha}$ denotes the $(1-\alpha)$th percentile of the standard normal distribution. Therefore, the required total sample size per group, N_{FEM}, for detecting an expected treatment effect $v = \Delta$ at the significance level α and with power $1-\beta$ for the MRCT under FEM would be $N_{\text{FEM}} = 2(z_{1-\alpha} + z_{1-\beta})^2\sigma^2/\Delta^2$.

10.2.2 Continuous Random Effects Model

In practice, regional variability caused by differences in ethnicity, environment, culture, and medical practice has been observed and may have impact upon a medicine's effect. An example regarding the possible impact of ethnic factors on the responses to therapeutics is the epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor gefitinib (Iressa). The Iressa trials have revealed significant variability in the response to gefitinib, with higher responses observed in Japanese patients than in a predominantly European population (27.5% vs. 10.4%) (Fukuoka et al. 2003).

Some insightful articles, such as Hung et al. (2010) and Wang and Hung (2012), have discussed regional heterogeneity and raised concerns that the assumption of a homogeneous treatment effect across regions may not be appropriate for MRCTs. Thus, many statisticians applied random effects model, which was originally proposed by DerSimonian and Laird for meta-analysis (DerSimonian and Laird 1986) , to MRCTs and assumed that regional treatment effects are random sample from a normal distribution (Chen et al. 2012; Quan et al. 2010). Here, this random effects model in MRCTs is denoted as the continuous random effect model (CREM). Many researchers intended to use CREM for solving the problem of heterogeneous treatment effects across regions. However, CREM assume that all regional treatment effects are unconditionally equal. Fundamentally, there is no difference between CREM and FEM. Therefore, CREM may be inappropriate for MRCTs when regional heterogeneity is considered (Please see Sect. 10.4.1. for the details).

Under CREM, regional treatment effects are assumed as a random sample from a normal distribution. That is,

$$\hat{v}_i | v_i \sim N\left(v_i, 2\sigma^2/N_i\right) \text{ and } v_i \sim N\left(v, \tau^2\right), \tag{10.3}$$

where $N(\mu, \zeta^2)$ represents a normal distribution with mean μ and variance ζ^2, and N_i is sample size per group in region i, $i = 1, \ldots, M$. Conditionally, as an estimate of

v_i, $\hat{v}_i$ is normally distributed with mean v_i and variance $2\sigma^2/N_i$. Unconditionally, as an estimate of v_i, $\hat{v}_i$ is normally distributed with mean v_i and variance $2\sigma^2/N_i + \tau^2$.

Under CREM, the overall treatment effect is a weighted average of each regional treatment effects, $v_{CREM} = \sum_{i=1}^{M} r_i v_i$, where $r_i = [\text{var}(v_i)]^{-1} / \sum_{i=1}^{M} [\text{var}(v_i)]^{-1}$. It is estimated by $\hat{v}_{\text{CREM}} = \sum_{i=1}^{M} \hat{r}_i \hat{v}_i$. The weights $\hat{r}_i = \left[\text{var}\left(\hat{v}_i\right)\right]^{-1} / \sum_{i=1}^{M} \left[\text{var}\left(\hat{v}_i\right)\right]^{-1}$, where $var(\hat{v}_i) = 2\sigma^2/N_i + \tau^2$. Thus, the weights $\hat{r}_i$ are proportional to the reciprocal sample variance of the regional means. Under CREM, the test statistic for the above hypothesis is

$$Z_{\text{CREM}} = \frac{\hat{v}_{\text{CREM}}}{\sqrt{\text{var}\left(\hat{v}_{\text{CREM}}\right)}}, \tag{10.4}$$

where the variance of $\hat{v}_{\text{CREM}}$ is $\text{var}\left(\hat{v}_{\text{CREM}}\right) = 1/\sum_{i=1}^{M} \left[\text{var}\left(\hat{v}_i\right)\right]^{-1}$. The null hypothesis would be rejected at the significance level α if the test statistic $Z_{\text{CREM}} > z_{1-\alpha}$, where $z_{1-\alpha}$ denotes the $(1-\alpha)$th percentile of the standard normal distribution. Furthermore, the required sample size per arm, N_{CREM}, for detecting an expected treatment effect $v = \Delta$ at the significance level α and with power $1-\beta$ for the MRCT under CREM is obtained by solving the following equation

$$\left(\frac{\Delta}{z_{1-\alpha} + z_{1-\beta}}\right)^2 = \frac{1}{\sum_{i=1}^{M} \left(\frac{2\sigma^2}{N_i} + \tau^2\right)^{-1}}, \tag{10.5}$$

Consider that $\hat{v}_{\text{CREM}}$ is asymptotically unbiased for v, with variance approximately equal to $1/\sum_{i=1}^{M} \left[\text{var}\left(\hat{v}_i\right)\right]^{-1}$, Chen et al. (2012) modified the test statistic with a t distribution when the number of regions is small. It is taken as $T_{\text{CREM}} = \hat{v}/\sqrt{S/(M-1)} \sim t_{M-1}$ with $S = \sum_{i=1}^{M} \hat{r}_i \left(\hat{v}_i - \hat{v}_{\text{CREM}}\right)^2 / \sum_{i=1}^{M} \hat{r}_i$ under H_0, where t_n represents the t distribution with degrees of freedom n. The total sample size required per group would be derived based on a non-central t distribution under the alternative hypothesis.

10.2.3 Discrete Random Effects Model (DREM)

Recognizing that regional treatment differences are typically not random samples from a normal distribution, Lan and Pinheiro (2012) proposed a discrete random effects model (DREM) to account for between-region variability for continuous responses. Lan et al. (2014) further applied DREM to time-to-event and binary responses. In this section, we introduce the discrete random effects model for continuous responses. Suppose that the patient population is divided into disjoint clinical regions S_1, S_2,…, S_M in an MRCT. The probability of a patient being randomly assigned to the ith region in the trial is $P(S_i) = W_i$, where $\sum_{i=1}^{M} W_i = 1$, for $i = 1$,

2,…, M. Theoretically, the sample sizes from the s regions are random, but in practice they can be replaced by the observed values, i.e., $w_i = N_i/N$.

At ith region, the treatment effect is a fixed number $v_i = \mu_{Ti} - \mu_{Ci}$. Let F be the treatment. Therefore, the treatment effect follows a discrete distribute $P\{F = v_i\} = W_i$. The overall treatment effect, v, is defined as the weighted sum of the regional treatment effects as $v_{\text{DREM}} = \sum_{i=1}^{M} W_i v_i$. The overall treatment effect is estimated by $\hat{v}_{\text{DREM}} = \sum_{i=1}^{M} W_i \hat{v}_i$. Correspondingly, the overall within-region variation and the between-region variation, are $\sum_{i=1}^{M} W_i 2\sigma^2 = 2\sigma^2$ and $\sum_{i=1}^{M} W_i (v_i - v)^2$, respectively.

Considering the same hypothesis in Eq. (10.1), the test statistics is given by

$$Z_{\text{DREM}} = \frac{\hat{v}_{\text{DREM}}}{\sqrt{\text{var}(\hat{v}_{\text{DREM}})}} = \frac{\hat{v}_{\text{DREM}}}{\sqrt{(2\sigma^2 + \tau^2)/N}} \tag{10.6}$$

under H_0. The test statistic Z_{DREM} is approximately normally distributed with a reasonably large sample size. The null hypothesis H_0 is rejected at the significance level α and the treatment T is claimed beneficial if the test statistic $Z_{\text{DREM}} > z_{1-\alpha}$, where $z_{1-\alpha}$ denotes the $(1 - \alpha)$th percentile of the standard normal distribution. Therefore, the power function for benefit is given by

$$PB = P[\text{Benefit}] = P(Z \geq z_{1-\alpha} | N, v) = \Phi\left(\frac{v}{\sqrt{(2\sigma^2 + \tau^2)/N}} - z_{1-\alpha} \right), \tag{10.7}$$

where Φ denotes the cumulative probability function of the standard normal distribution and $z_{1-\alpha} \approx 1.96$ if one-sided $\alpha = 0.025$.

Under DREM, the total required sample size N_{DREM} is planned for detecting an expected treatment difference $v = \Delta > 0$ at significance level α and power $1 - \beta$, satisfying

$$\left(\frac{\Delta}{z_{1-\beta} + z_{1-\alpha}} \right)^2 = \frac{2\sigma^2 + \tau^2}{N_{\text{DREM}}}. \tag{10.8}$$

However, the overall treatment difference Δ is difficult to pre-specify when planning an MRCT. In practice, all regional treatment effects $\{v_i\}$ are unknown and hard to pre-specify at design stage. The expected treatment difference Δ cannot be pre-assigned. Lan et al. (2014) provided a suggestion to address this issue by assuming that all the possible values of effects $\{v_i\}$ fall into an interval $[A, B]$. An over-estimated τ^2 can be obtained by $\tilde{\tau}^2 = (B - A)^2/4$. A possible Δ may be chosen by $(A + B)/2$. Giving values of v, σ^2, using $\tilde{\tau}^2$, a conservative required sample size N_{DREM} can be acquired.

The right side of Eq. (10.8) is much simpler than that of Eq. (10.5). In other words, the sample size determination under CREM is more complicated than that

under DREM. This is the difficulty of designing an MRCT in practice by using CREM. Thus, the required sample size can be determined by

$$N_{\text{DREM}} = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{\Delta}\right)^2 \left(2\sigma^2 + \tau^2\right). \tag{10.9}$$

10.3 Consistency and Inconsistency

The primary objective of an MRCT is to demonstrate the overall treatment effect based on data from the whole trial. As we mentioned at the beginning of this chapter, the underlying assumption of MRCTs is the consistency of the treatment effects across regions so the trial results are applicable to each region. However, it not easy to define a criterion to confirm a consistent trend of treatment effects in all regions. In 2007, the Japanese Ministry of Health, Labour, and Welfare (MHLW) published the "Basic Principles on Global Clinical Trials" guidance for planning and implementation of global clinical studies. It focuses on how to assess the efficacy of a drug in all participating regions (the "probability of benefit" introduced in Sect. 10.2.3) and how to evaluate the possibility of applying the overall trial results to each region (the "probability of benefit and consistency") by conducting an MRCT (Ministry of Health, Labour and Welfare of Japan (MHLW) 2007). The Japanese MHLW guideline provided two criteria for establishing the efficacy in a specific region and the consistency in efficacy among regions. Let D be the estimated overall treatment difference of all participants, and D_i be the treatment effect of the ith region, respectively; $i = 1, 2, \ldots, M$.

Method 1 (M1): $D_i/D > \pi$ (with $\pi \geq 0.50$) for a specific region i.
Method 2 (M2): $D_i > 0$ for all i.

Method 1 illustrates the consistency between the results of the Japanese region and the overall result. If the ratio of the treatment effect estimate of the Japanese region to that of the overall regions is greater than 0.5, the consistent trend of treatment effects across regions is confirmed. Method 2 assesses the consistency among all participating regions in an MRCT. It describes consistency in the sense that in addition to overall treatment in all regions combined, the estimate of the treatment effect in each region needs to exceed zero.

A significant overall result in an MRCT would likely be regarded as a successful global trial. However, local regulatory authorities may want to confirm the consistent trend of the treatment effect in the region under their administration. If the "probability of benefit and consistency (*PBC*)" is considered at the planning stage, the *PBC* result can be directly applied in the MRCT. *PBC* may be a high hurdle for MRCTs. Therefore, at the planning stage, regional differences that would cause inconsistency should be examined via genetic and epidemiology research, and regional heterogeneity research. Regions that are potentially inconsistent should not be included in the MRCT.

Many research articles have discussed approaches to assess a consistent trend of treatment effects considering methods 1 and 2. Kawai et al. (2008) proposed an approach to assess the consistency criterion by assuming a uniform treatment effect across regions; and Tanaka et al. (2012) focused on qualitative consistency. Both papers are based on method 2 of the MHLW. Other articles have attempted to describe approaches to assess the consistency of treatment effects based on method 1 of the MHLW, such as Chen et al. (2012), Quan et al. (2010), Uesaka (2009), Chen et al. (2010), Ko et al. (2010), Tsou et al. (2010, 2011, 2012), and Quan et al. (2010, 2013). Next, we assume that treatment effects follow the DREM, and extend the consistency consideration based on method 2 of the MHLW to construct the "probability of benefit and consistency."

*10.3.1 Probability for Consistency (*PC*) Under DREM*

In Sect. 10.2.3, we introduced the power for benefit (*PB*). Here, we extend this to the power for benefit and consistency. Let us define the following notation.

$$PC = P[\text{Consistency (M2)}|N, \theta] = P(\hat{v}_i > 0, \text{ for all } i|N, v),$$

where $\theta = (\sigma^2, \tau^2, v_1, \ldots, v_M, W_1, \ldots, W_M)$; W_i denotes the proportion of patients; $i = 1, 2, \ldots, M$; and $\sum W_i = 1$.

Kawai et al. (2008) derived the probability of consistency based on method 2 under the uniform treatment effect across regions as follows.

$$PC_{\text{FIX}} = P[\hat{v}_1 > 0, \ldots, \hat{v}_M > 0] = \prod_{i=1}^{M} \Phi\left(\sqrt{W_i} \cdot \left(z_{1-\alpha} + z_{1-\beta}\right)\right)$$

Here, PC_{FIX} depends only on the number of regions M and the proportion of patients W_i in all regions.

Under DREM, the power of the consistency based on method 2 can be derived as follows. For the ith region, $i = 1,\ldots, M$, the test statistic is

$$Z_i = \frac{\hat{v}_i}{\sqrt{\left(2\sigma^2 + \tau^2\right)/N W_i}},$$

and $Z_{\min}$ denotes the minimum of the M statistics, $Z_{\min} = \min\{Z_1, Z_2, \ldots, Z_M\}$. Therefore, the power for consistency (PC) under DREM is given by

$$PC = P[\text{Consistency (M2)}|N, \theta] = P(\hat{v}_i > 0, \text{ for all } i|N, \theta) = P(Z_{\min} > 0)$$

$$= \prod_{i=1}^{M} \left\{ 1 - \Phi\left(\frac{-v_i}{\sqrt{\frac{2\sigma^2+\tau^2}{N_i}}} \right) \right\} = \prod_{i=1}^{M} \left\{ \Phi\left(\frac{v_i}{\sqrt{\frac{2\sigma^2+\tau^2}{N \cdot W_i}}} \right) \right\}, \tag{10.10}$$

where $Z = \sum_{i=1}^{M} \sqrt{W_i} Z_i$ and Φ denotes the cumulative probability function of the standard normal distribution. If we replace N as N_{DREM} in Eq. (10.9), the PC in Eq. (10.10) can be rewritten as

$$PC = \prod_{i=1}^{M} \Phi\left(\frac{\sqrt{W_i} \cdot v_i}{\Delta} \cdot \left(z_{1-\alpha} + z_{1-\beta} \right) \right). \tag{10.11}$$

Under DREM, Eqs. (10.10) and (10.11) show that PC depends on the number of regions, sample size proportions $\{W_i\}$, and treatment effects $\{v_i\}$ in all regions, within-group variance σ^2, and between-region variance τ^2. Moreover, PC is an increasing function of the effect size $v_i/(\sigma^2 + \tau^2)$ for a fixed N.

10.3.2 Optimal Allocation Among Regions to Maximize PC

When planning an MRCT, the weights W_i should be determined for efficientively demonstrating the efficacy and consistency of the test drug. The determination of weights $\{W_i, i=1,\ldots, M\}$ is a better approach than relying on a decision-maker's preferences. By maximizing power for consistency (PC) under DREM, an optimal allocation of patients among regions can be obtained. Liu et al. (2016) showed that PC under DREM is maximized when $\sqrt{W_1} \cdot v_1 = \cdots = \sqrt{W_M} \cdot v_M$ (Liu et al. 2016). As we have seen, the optimal allocation $\{W_i\}$ only depends on the values of $\{\sqrt{W_i} \cdot v_i\}$ under DREM. A special case is that all regional effects are equal and PC is maximized when $W_1 = \cdots = W_M = 1/M$, which is reduced to the finding in Kawai et al. (2008).

*10.3.3 Probability for Benefit and Consistency (*PBC*)*

Now we consider a more complex issue on consistency assessment, which is the evaluation of probability for benefit and consistency (PBC). Obviously, this is more difficult than considering only probability for benefit (PB) or the probability for consistency (PC). In this section, we define the PBC using M2 consistent criterion of MHLW (2007).

For a fixed N, the power for benefit and consistency (PBC) is denoted by $PBC(N) = \mathrm{P}[Z > z_{1-\alpha} \ \& \ Z_{\min} > 0 \mid N]$. Let parameter space $\Theta = \{\text{All } \theta = (W_i, v_i, i = 1, 2,\ldots, M) \text{ under consideration}\}$. The theoretical formula of $PBC(N, \theta)$ is derived as

$$PBC(N,\theta)=\left\{\int_{c_M}^{\infty}\cdots\int_{c_2}^{\infty}\left[\int_{\max\left\{c_1,\frac{A-\left(\sum_{k=2}^{M}d_k u_k\right)}{d_1}\right\}}^{\infty}\phi(u_1)du_1\right]\phi(u_2)du_2\cdots\phi(u_M)du_M\right\} \tag{10.12}$$

where $c_i=\frac{-v_i}{\sqrt{(2\sigma^2+\tau^2)/N_i}}$, $d_i=\sqrt{W_i}$, $A=z_{1-\alpha}-\frac{v}{\sqrt{(2\sigma^2+\tau^2)/N}}$, and $i=1, 2, \ldots, M$ (Liu et al. 2016). The theoretical formula of $PBC(N, \theta)$ shows that PBC increases with N for any fixed θ. Liu et al. (2016) pointed out that $PBC(N, \theta)=PC(N, \theta)\times P[Z>z_{1-\alpha}\mid N, \theta, Z_{\min}>0]$, where $P[Z>z_{1-\alpha}\mid N, \theta, Z_{\min}>0]$ indicates the statistical power conditioning on $Z_{\min}>0$. On the other hand, when PBC and PC are known the conditional power $P[Z>z_{1-\alpha}\mid N, \theta, Z_{\min}>0]$ could be calculated. Table 10.1 shows the PBC and PC, along with the conditional power $P[Z>z_{1-\alpha}\mid N, \theta, Z_{\min}>0]$ under DREM for various values of θ, including theoretical and simulated values. Simulated values are in brackets with 100,000 simulation times. The simulated values are very close to the theoretical values in this case. Moreover, we can see that the conditional power $P[Z>z_{1-\alpha}\mid N, \theta, Z_{\min}>0]$ is not a constant over Θ and that PBC increases with PC.

Table 10.1 Powers $PBC(N, \theta)$ and $PC(N, \theta)$, and conditional power $P[Z>z_{1-\alpha}\mid N, \theta, Z_{\min}>0]$

W_1	W_2	W_3	ν_3	τ^2	$PC(N, \theta)$	$PBC(N, \theta)$	$P[Z>z_{1-\alpha}\mid N, \theta, Z_{\min}>0]$
0.1	0.1	0.8	0.22	0.003	0.5055 (0.5112)	0.4784 (0.4868)	0.9465 (0.9523)
0.15	0.15	0.7	0.23	0.004	0.5537 (0.5592)	0.5279 (0.5357)	0.9533 (0.9580)
0.2	0.2	0.6	0.25	0.006	0.5899 (0.5946)	0.5653 (0.5716)	0.9584 (0.9614)
0.25	0.25	0.5	0.28	0.008	0.6179 (0.6230)	0.5948 (0.6023)	0.9625 (0.9668)
0.3	0.3	0.4	0.31	0.012	0.6413 (0.6471)	0.6195 (0.6281)	0.9660 (0.9706)

Let $\nu_1=0.05$, $\nu_2=\nu=0.2$, $\alpha=0.025$, $PB=1-\beta=0.90$, desired PBC level $=\gamma=0.85$ for $W_1=W_2<W_3$. Theoretical values of $PC(N, \theta)$ and $PBC(N, \theta)$ were calculated by Eqs. (10.10) and (10.12), respectively. Values in parentheses are the corresponding simulated probabilities with simulation times = 100,000

10.3.4 Iteration Procedures to Derive $\mathbf{N}^*$ for a Target Level of PBC

At the beginning of the design stage, we considered the power of benefit to construct an MRCT and the required sample size N is calculated by satisfying $PB(N, \theta) \geqq 1 - \beta$ for given parameter θ. Then, a sample size N^* is derived considering the benefit and consistency simultaneously, such as $PBC(N^*, \theta) \geqq \gamma$, where γ is a desired level (e.g., 85% or 90%). If $PBC(N, \theta) \geq \gamma$, then $N^* = N$ is the desired sample size. If $PBC(N, \theta) < \gamma$, then we need to find a larger sample size $N^* > N$ so that $PBC(N^*, \theta) \geq \gamma$. Liu et al. provided three different algorithms for deriving sample size at the desired level of power for benefit and consistency (Liu et al. 2016). Here, we introduce the most efficient one.

Efficient algorithm

We apply the relationship between *PBC* and sample size under DREM as an algorithm for searching N^* at the desired level γ, yielding

$$\frac{N^*}{N_0} = \left(\frac{z_{1-\alpha} + z_{\gamma}}{z_{1-\alpha} + z_{\gamma_0}} \right)^2. \tag{10.13}$$

This is the iteration procedure we adapted for the derivation of N^*. From $PB = 1 - \beta = \gamma_0$, we find initial sample size N_0. If $PBC(N_0, \theta) \geqq \gamma_0$, then $N^* = N_0$. Otherwise, replace N_0 with the value of N^* derived from Eq. (10.13). Repeat the procedure until $PBC(N^*, \theta) \geqq \gamma$.

10.4 A Comparison of CREM and DREM

10.4.1 Problems with CREM

CREM assumes that heterogeneous treatment effects are random and follow a normal distribution with mean v and between-region variability τ^2. CREM assumes that all v_i's are unconditionally equal. Fundamentally, there is no difference between CREM's and FEM's framework. To properly estimate τ^2, the number of regions should not be too small or region should be defined at the country level, according to (Quan et al. 2017). However, when the number of regions is greater than 5, the consistency may not be properly assessed. This placed CREM in a dilemma. In addition, the assumption that the regional treatment effects are random samples from a normal distribution is questionable.

Another point is that the variances of the treatment effect estimates in DREM and CREM are completely different. Under DREM and CREM, $Var\left(\hat{v}_{DREM}\right) = (2\sigma^2 + \tau^2)/N$ and $Var\left(\hat{v}_{CREM}\right) = 1/\Sigma_i(2\sigma^2/N_i + \tau^2)$. It should be noticed that the parameter being estimated by $Var\left(\hat{v}_{CREM}\right)$ is not the mean of the discrete distribution F, but

rather a linear combination of the regional means v_1, v_2, …, v_M (with coefficients depending on the regional sample sizes and the variance components). When taken as an estimate of v_{CREM}, $\hat{v}_{CREM}$ is a biased estimator, except on all regions have the same mean, i.e., the fixed effects model. We use an example with a discrete prior to illustrate this. Suppose that we have an MRCT with 600 subjects per arm (total sample size = 1200) in 3 regions. We assume that the within-region variations are identical for all regions and equal to 25, i.e. $\sigma^2 = 25$. The following table gives further assumptions on the underlying parameters and distributions.

i	N_i	W_i	v_i
1	120	0.2	8
2	240	0.4	10
3	240	0.4	12

It follows that the overall treatment effect is v_{DREM}= 10.4 and the between-region variation is $\tau^2 = 2.24$. Then, the variance of $\hat{v}_{DREM}$ is $Var(\hat{v}_{DREM}) = 0.0908$. Under CREM, the variance of sample means would be 2.66, 2.45, and 2.45, giving the weights of CREM as r_i=(0.3154, 0.3423, 0.3423). For CREM, the overall mean, v_{CREM}, is 10.05 and the corresponding variance is $Var(\hat{v}_{CREM}) = 0.8380$. In this case, v_{CREM} is closed to v_{DREM}, but biased. Hung et al. (2010) mention that the use of relative sample size weights will provide an unbiased estimator for the population mean. In addition, the variance of overall mean under CREM is higher than that under DREM. It means that DREM provides an estimate of the overall treatment effect with more appropriate precision.

Other possible problems of CREM are simply listed as follows.

(a) The overall mean in CREM may not be easily interpreted or defined in general cases, especially to those who lack training in statistics, because of the complicated weight structure.
(b) When conducting an MRCT, it is possible that some regions join the MRCT one or two years later or some regions are combined at the end of the trial. Under CREM, all weights $\{W_i\}$ are pushed to $1/M$ as regional sample size N_i tend to infinity. In such cases, the weights at the end of a trial could be very different from the pre-determined weights. This violates the one-patient-one-vote principle.
(c) The power for sample size determination under CREM will never reach 1 when regional sample size N_i tends infinity.

10.4.2 Comparison Between CREM and DREM: Sample Size Determination

In practice, the sample size determination plays an important role in implementation of a clinical trial. For discovering the difficulty in application of CREM, we compare

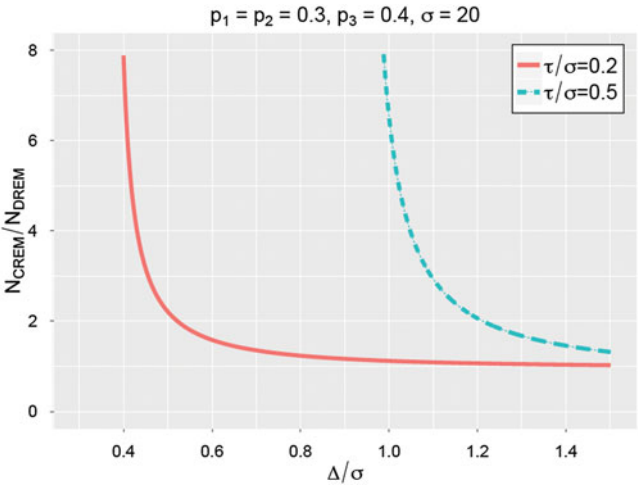


Fig. 10.1 Sample-size ratio $N_{\text{CREM}}/N_{\text{DREM}}$ versus Δ/σ at $\tau/\sigma = 0.2$ and 0.5

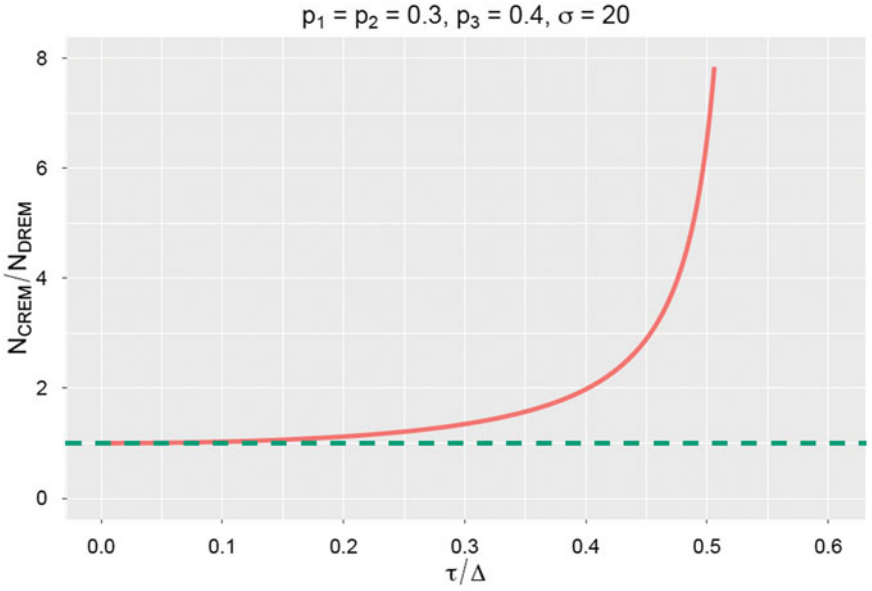


Fig. 10.2 Sample-size ratio $N_{\text{CREM}}/N_{\text{DREM}}$ versus τ/Δ

the required sample size in CREM with that in DREM. We use a numerical example to compare sample sizes in CREM and DREM. Assume an MRCT with three regions and given $p_1 = p_2 = 0.3$, $p_3 = 0.4$, $\sigma = 20$, $\alpha = 0.05$, and $1 - \beta = 0.9$.

Figure 10.1 displays the sample-size ratio as a function of effect size Δ/σ for CREM compared with DREM, for ratio $\tau/\sigma = 0.2$ and 0.5. It is clear that the

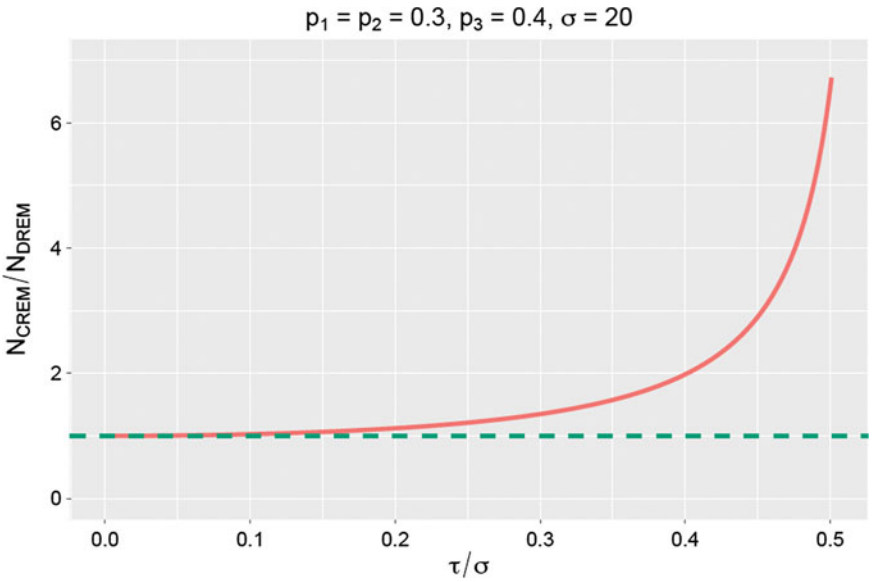


Fig. 10.3 Sample-size ratio $N_{\text{CREM}}/N_{\text{DREM}}$ versus τ/σ

ratio (N_{CREM}/N_{DREM}) decreases with increasing value of Δ/σ and increases with increasing value of τ/σ. The ratio (N_{CREM}/N_{DREM}) is always higher than 1 when the between-region variation exists. Moreover, CREM needs a larger sample size than DREM when the between-region variance is large relative to the within-region variance, such as $\tau/\sigma = 0.5$.

Figure 10.2 shows the sample-size ratio as a function of effect size Δ/σ for CREM compared with DREM. Given fixed value δ, the ratio $N_{\text{CREM}}/N_{\text{DREM}}$ increases as the ratio τ/Δ, while the values of between-region variance τ are increased. In Fig. 10.2, the sample-size ratio $N_{\text{CREM}}/N_{\text{DREM}}$ is always larger than 1. The required sample size in CREM increases much faster than the sample size in DREM with high value of between-region variance τ. For example, N_{CREM} is almost eight times larger than N_{DREM} when $\tau/\Delta > 0.5$.

For $0 \leq \tau/\sigma \leq 0.5$, and given $\Delta = 20$, $\sigma = 20$, $p_1 = p_2 = 0.3$, $p_3 = 0.4$, $\sigma = 20$, $\alpha = 0.05$, and $1 - \beta = 0.9$, the plots of sample-size ratios of CREM and DREM against values of τ/σ are given in Fig. 10.3. As shown in that figure, N_{CREM} increases greatly compared with N_{DREM} as between-region variance increases. For instance, for $\tau/\sigma = 0.5$ the sample size of CREM reaches 6.7-fold compared with that of DREM. Therefore, high heterogeneity among regions requires higher samples sizes with CREM than with DREM; this should be taken into account when planning an MRCT considering regional difference.

As explored in the numerical example, using CREM to estimate sample size may be inappropriate for designing an MRCT and make it difficult to implement an MRCT, especially when the between-region variance τ^2 is relatively large. Sample

size estimated under DREM (N_{DREM}) may be more applicable for detecting the overall treatment effect $\Delta > 0$.

10.5 Drop-Min Data Analysis

Inconsistency among regions in multi-regional clinical trials (MRCT) is noted in some drug development programs. The inconsistency may be due to differences in extrinsic factors; for example, the clopidogrel trial COMMIT (EMA 2015) was intended to evaluate clopidogrel plus aspirin in comparison with placebo plus aspirin in treatment of patients with acute myocardial infarction but the use of beta-blockers caused concerns for the Committee for Medicinal Products for Human Use (CHMP) regarding relevance of the EU clinical setting because most European patients got beta blockers. Results thus might not be relevant to European population. Therefore, CHMP considered the COMMIT trial as supportive rather than pivotal. The inconsistency may also be due to intrinsic factors. In the gefitinib development program, survival effect was only seen in Asian patients which was probably due to different tumor genetics among Asian and European patients. As a result, gefitinib was only approved in Japan not in the European Union (Dunder 2009).

Inconsistency may be defined in many different ways. Regardless of the use of the definitions, the inconsistent regions are the ones with minimum observed treatment effect. Therefore, the "drop-min"' approach is to estimate the treatment effects when the region with minimum observed treatment effect is excluded from data analysis. Similar to selecting the winner as a screening process, dropping the loser naturally introduces bias to naïve statistical inference for the remaining subgroups, centers or regions.

For an MRCT with M regions, assume that the treatment effect in ith region is X_i, for $i = 1, 2, \ldots, M$, and X_i follows a normal distribution. Let $w_i = \frac{N_i}{N}$. Denote that $X_{(i)}$ is the order statistics of X_i, i.e., $X_{(1)} < X_{(2)} < \ldots < X_{(M)}$. If the minimum is inconsistent to $\{X_{(2)}, \ldots, X_{(M)}\}$ and be dropped, the over treatment result would be estimated by $\{X_{(2)}, \ldots, X_{(M)}\}$. However, the estimate of the treatment effect is a biased estimator based on the remaining regions $\{X_{(2)}, \ldots, X_{(M)}\}$, because $X_{(1)}$ is dropped. For example, $M = 2$, let $X_i \sim N(0, 1)$, for $i = 1, 2$. Define that $X_{(1)}$ is the minimum of $\{X_1, X_2\}$ and $X_{(2)}$ is the maximum of $\{X_1, X_2\}$. If we drop $X_{(1)}$ and the overall mean would be estimated by $X_{(2)}$. The expectation and variance of $X_{(2)}$ are $\mathrm{E}[X_{(2)}] = 0.5642$ and $\mathrm{Var}[X_{(2)}] = 0.6817$ $(1 - 0.5942^2)$, respectively. This example shows that $\mathrm{E}[X_{(2)}] \neq 0$ and $\mathrm{Var}[X_{(2)}] \neq 1$, the biased estimator and variance result in an inappropriate test statistics. For data analysis, the test statistics should be modified. Shun et al. (2008) shows that the adjusted Z statistic has a skew normal distribution. Next, we introduce the drop-min approach for the bias and variance calculation for FEM and DREM. CREM involves estimating τ^2, which may not be properly estimated due to a small number of regions in an MRCT. Therefore, we do not recommend for implementation although a similar drop-min approach applies to CREM.

Drop-min for FEM. Recall that the fixed effects model assumes $v_1 = \cdots = v_M = v$. The estimator of v with $X_{(1)}$ excluded is

$$\tilde{v}_- = \sum_{i=2}^{M} \frac{N}{N - N_{(i)}} w_{(i)} X_{(i)},$$

where $N_{(i)}$ and $w_{(i)}$ are the sample size and weight associated with $X_{(i)}$, respectively. Its bias is calculated as

$$B_M = E\big[\tilde{v}_- - v\big] = E\left[\sum_{i=2}^{M} \frac{N}{N - N_{(i)}} w_{(i)} \big(X_{(i)} - v\big)\right].$$

Denote $V_M = Var\big[\tilde{v}_-\big]$. The test statistic for the drop-min approach is

$$Z_M^* = \frac{\tilde{v}_- - B_M}{\sqrt{V_M}}.$$

Although the explicit formula for B_M and V_M are not available, they can be calculated via the resampling approach. The confidence intervals of v will also be constructed.

Suppose $\hat{\sigma}^2$ is an estimate of σ^2. In simulation, set $\sigma^2 = \hat{\sigma}^2$. The simulation takes two steps.

Step 1, generate S sets of random samples of $X_1 \sim N\Big(0, \frac{\sigma^2}{N_1}\Big), \ldots, X_M \sim N\big(0, \sigma^2/N_M\big)$. For each set of random sample, let $x_{(1)}, \ldots, x_{(M)}$ be the ordered $x_1, \ldots, x_M$. Let $N_{(i)}$ and $w_{(i)}$ are the sample size and weight associated with $x_{(i)}$, respectively. Calculate

$$b_M = \sum_{i=2}^{M} \frac{N}{N - N_{(i)}} w_{(i)} x_{(i)}.$$

Step 2, obtain the empirical distribution of these b_M's. Calculate its $(\alpha/2)^{th}$- and $(1 - \alpha/2)^{th}$-quantiles and denote them by $q_{\alpha/2}$ and $q_{1-\alpha/2}$, respectively. A 2-sided α-confidence interval of v is $\big(\tilde{v}_- + q_{\alpha/2}, \tilde{v}_- + q_{1-\alpha/2}\big)$. B_M and V_M are calculated the mean and variance of b_M's, respectively.

Drop-min for DREM. DREM is fundamentally different from the FEM. Rather than assuming a common treatment effect v, DREM treats the M regional effects $v_1, \ldots, v_M$ and their corresponding sample size $(N_1, \ldots, N_M)$ as population parameters. The overall treatment effect is the combined effect of all regions $= \tilde{v}_- = \sum_{i=1}^{M} w_i X_i$, where $w_i = \frac{N_i}{N}$. However, suppose X_m of region m is the observed minimum treatment effect, and is dropped from the analysis. This means that, under the DREM assumption, changing the components of an MRCT, e.g., the drop-minimum analysis, will change the combined effect definition. The combined effect among

the rest of regions $v_{-m} = \sum_{i=1, i\neq m}^{M} w_{im} v_i$, where $w_{im} = \frac{N_i}{N-N_m} = \frac{N}{N-N_m} w_i$. The combined effect is conditional one. Naturally, v_{-m} is estimated by

$$\tilde{v}_{-m} = I_{\{X_m = X_{(1)}\}} \sum_{i=1, i\neq m}^{M} w_{im} X_i,$$

Its bias is calculated as

$$B_{-m} = E\left[\tilde{v}_{-m} - v_{-m} | X_m = X_{(1)}\right] = \frac{N}{N-N_m} E\left[\sum_{i=1, i\neq m}^{M} w_i (X_i - v_i) | X_m = X_{(1)}\right].$$

Denote $V_{-m} = Var\left[\tilde{v}_{-m}\right]$. The test statistic for the drop-min approach is

$$Z^*_{-m} = \frac{\tilde{v}_{-m} - B_{-m}}{\sqrt{V_{-m}}}.$$

Again, the explicit formula for B_{-m} and V_{-m} are not available, they can be calculated via the resampling approach and the confidence intervals of v_{-m} will also be constructed.

Suppose $\hat{\sigma}^2$ is an estimate of σ^2. In simulation, set $\sigma^2 = \hat{\sigma}^2$. The observed regional treatment effects X_i will be used as the regional effects v_i and sample sizes N_i as population weights w_i, i.e., set $v_i = X_i$ and as population weights $w_i = N_i/N$, The simulations takes two steps.

Step 1, generate S sets of random samples of $X_1 \sim N\left(v_1, \frac{\sigma^2}{N_1}\right), \ldots, X_M \sim N\left(v_M, \sigma^2/N_M\right)$. For each set of random sample, if x_m is the minimum, calculate

$$b_{-m} = \frac{N}{N-N_m} \sum_{i=1, i\neq m}^{M} w_i (X_i - v_i).$$

Step 2, obtain the empirical distribution of these b_{-m}'s. Calculate its $(\alpha/2)^{th}$- and $(1-\alpha/2)^{th}$-quantiles and denote them by $q_{\alpha/2}$ and $q_{1-\alpha/2}$, respectively. A 2-sided α-confidence interval of v_{-m} is $\left(\tilde{v}_{-m} + q_{\alpha/2}, \tilde{v}_{-m} + q_{1-\alpha/2}\right)$. B_M and V_M are calculated the mean and variance of b_{-m}'s, respectively.

10.6 Discussion

In an MRCT, the primary objective is to demonstrate the overall treatment effect based on data from all participating regions. How to combine evidence of treatment effects from different regions is an important problem in MRCTs, especially when regional heterogeneity is marked. In this chapter, we review three models (FEM, CREM, and

DREM) for estimating overall treatment effect and interpreting the trial results. In addition, local regulatory authorities may request to evaluate the consistency among regions after the overall treatment effect is demonstrated in an MRCT. Thus, we explore the impact on sample size requirement by prospectively taking consistency among regions into account when designing an MRCT. We further compare CREM and DREM on different aspects (e.g., model assumption, weights for combining treatment effects among regions, one-patient-one vote principle) to understand the possible dilemma of regular random effects model. We also review a "drop-min" approach for analyzing data when a region with minimum observed treatment effect is excluded from data analysis.

There is no perfect model for combining treatment effects among regions. There is no best approach for the design, implement, analysis and interpretation of result of MRCTs, either. Through accumulating knowledge on the features of different approaches and experiences on conducting MRCTs, we may apply a more appropriate approach to improve the efficiency of an MRCT.

References

Chen, C. T., Hung, H. M. J., & Hsiao, C. F. (2012). Design and evaluation of multiregional trials with heterogeneous treatment effect across regions. *Journal of Biopharmaceutical Statistics, 22,* 1037–1050.

Chen, J., Quan, H., Binkowitz, B., Ouyang, S. P., Tanaka, Y., Li, G., Menjoge, S., Ibia, E. (2010). for the Consistency Workstream of the PhRMA MRCT key issue team. Assessing consistent treatment effect in a multi-regional clinical trial: A systematic review. *Pharmaceutical Statistics, 9,* 242–253. https://doi.org/10.1002/pst.438.

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

Dunder, K. (2009). Issues on acceptance of foreign clinical data in Europe from a regulatory perspective 2009. In 9th Kitasato University-Harvard School of Public Health Symposium.

EMA, Epar, scientific discussion, variation application for clopidogrel 2015.

Fukuoka, M., Yano, S., Giaccone, G., Tamura, T., Nakagawa, K., Douillard, J. Y., et al. (2003). Multi-institutional randomized phase II trial of gefitinib for previously treated patients with advanced non-small-cell lung cancer (The IDEAL 1 Trial). *Journal of Clinical Oncology, 21,* 2237–2246.

Hung, H. M. J., Wang, S. J., & O'Neill, R. T. (2010). Consideration of regional difference in design and analysis of multi-regional trials. *Pharmaceutical Statistics, 24,* 173–178. https://doi.org/10.1002/pst.440.

International Concil for Harmonization of Technical Requirement for Pharmaceuticals for Human Use, ICH E17. (2016). General Principles for Planning and Design of Multi-regional Clinical Trials (draft).

Kawai, N., Stein, C., Komiyama, O., & Li, Y. (2008). An approach to rationalize partitioning sample size into individual regions in a multiregional trial. *Drug Information Journal, 42,* 139–147.

Ko, F. S., Tsou, H. H., Liu, J. P., & Hsiao, C. F. (2010). Sample size determination for a specific region in a multi-regional trial. *Journal of Biopharmaceutical Statistics, 24,* 870–885.

Lan, K. K. G., & Pinheiro, J. (2012). Combined estimation of treatment effects under a discrete random effects model. *Statistics in Biosciences, 4,* 235–244.

Lan, K. K. G., Pinheiro, J., & Chen, F. (2014). Designing multiregional trials under the discrete random effects model. *Journal of Biopharmaceutical Statistics, 24,* 415–428.

Liu, J. T., Tsou, H. H., Lan Gordon, K. K., Chen, C. T., Lai, Y. H., Chang, W. J., Tzeng, C. S., Hsiao, C. F. (2016, June). Assessing the consistency of the treatment effect under the discrete random effects model in multiregional clinical trials. *Statistics in Medicine, 35*(14), 2301–2314.

Ministry of Health, Labour and Welfare of Japan (MHLW). Basic principles on global clinical trials 2007. Available at http://www.pmda.go.jp/files/000153265.pdf. Accessed date May 4, 2017.

Quan, H., Li, M., Chen, J., Gallo, P., Binkowitz, B., Lbia, E., et al. (2010b). Assessment of consistency of treatment effects in multiregional clinical trials. *Drug Information Journal, 44,* 617–632.

Quan, H., Li, M., Shih, W. J., Ouyang, S. P., Chen, J., Zhang, J., et al. (2013). Empirical shrinkage estimator for consistency assessment of treatment effects in multi-regional clinical trials. *Statistics in Medicine, 32,* 1691–1706. https://doi.org/10.1002/sim.5543.

Quan, H., Mao, X., Tanaka, Y., Binkowitz, B., Li, G., Chen, J., et al. (2017). Example-based illustrations of design, conduct, analysis and result interpretation of multi-regional clinical trials. *Contemporary Clinical Trials, 58,* 13–22.

Quan, H., Zhao, P. L., Zhang, J., Roessner, M., & Aizawa, K. (2010a). Sample size considerations for Japanese patients in a multi-regional trial based on MHLW Guidance. *Pharmaceutical Statistics, 9,* 100–112. https://doi.org/10.1002/pst.380.

Shun, Z., Lan, K. K. G., Soo, Y. (2008). Interim treatment selection using the normal approximation approach in clinical trials. *Statistics in Medicine, 27*(4).

Tanaka, Y., Li, G., Wang, Y., & Chen, J. (2012). Qualitative consistency of treatment effects in multiregional clinical trials. *Journal of Biopharmaceutical Statistics, 22,* 988–1000.

Tsou, H. H., Chien, T. Y., Liu, J. P., & Hsiao, C. F. (2011). A consistency approach to evaluation of bridging studies and multiregional trials. *Statistics in Medicine, 30,* 2171–2186.

Tsou, H. H., Chow, S. C., Lan, K. K. G., Liu, J. P., Wang, M., Chen, H. D., et al. (2010). Proposals of statistical consideration to evaluation of results for a specific region in multi-regional trials—Asian Perspective. *Pharmaceutical Statistics, 9,* 201–206. https://doi.org/10.1002/pst.442.

Tsou, H. H., Hung, H. M. J., Chen, Y. M., Huang, W. S., Chang, W. J., & Hsiao, C. F. (2012). Establishing consistency across all regions in a multi-regional clinical trial. *Pharmaceutical Statistics, 11,* 295–299. https://doi.org/10.1002/pst.1512.

Uesaka, H. (2009). Sample size allocation to regions in a multiregional trial. *Journal of Biopharmaceutical Statistics, 19,* 580–594.

Wang, S. J., & Hung, H. M. J. (2012). Ethnic sensitive or molecular sensitive beyond all regions being equal in multiregional clinical trials. *Journal of Biopharmaceutical Statistics, 22,* 879–893.

Chapter 11
Multi-Regional Clinical Trials, ICH-E17, and Subpopulations

Yoko Tanaka, Bruce Binkowitz and Bill Wang

11.1 Introduction

Drug development has rapidly been globalized, and multi-regional clinical trials (MRCTs) have widely been conducted for the purpose of regulatory filing in multiple regions using the same trial data in inside and outside of ICH (international council of harmonization) regions (Regions considered as ICH are European Union, US, Japan, Canada, and Switzerland). Regulatory agencies face challenges in evaluating data from MRCTs for drug approval. Although the Q&A of the ICH E5 guideline (ICH E5 1998, 2006) partially covers issues relating to MRCTs, there was no harmonized ICH Guideline on MRCTs, especially focusing on scientific issues in planning/designing MRCTs. A lack of harmonization on this topic may cause additional burden for sponsor and difficult situation for conducting MRCTs (ICH E17 2014). Therefore, an expert working group (EWG) was established in 2014 to develop a new guideline, ICH-E17 (2016, General principles for planning and design of multi-regional clinical trials). ICH-E17 EWG members and observers consisted of regulatory/Industry membership from EU, Japan, US, Canada, WHO, GCC (Saudi), Brazil, Singapore, Taiwan, and Korea.

The main objective of ICH-E17 (2016) is to provide common points to consider in planning/designing MRCTs and minimize conflicting opinions from regulatory bod-

Y. Tanaka (✉)
Santen Inc., 6401 Hollis St., Emeryville, CA 94608, USA
e-mail: yoko.tanaka@santen.com

B. Binkowitz
Biometrics, Shionogi, Inc., 300 Campus Dr, Florham Park, NJ 07932, USA
e-mail: bruce.binkowitz@gmail.com

B. Wang
Biostatistics and Research Decision Sciences, Merck Research Laboratories, Merck & Co., Inc, Kenilworth, NJ, USA
e-mail: william_wang@merck.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_11

ies, hence increasing the acceptability of MRCTs in global regulatory submissions. As discussed in the draft ICH-E17 guidance (2016), ethnic factors are a major point of consideration when planning MRCTs. They should be identified during the planning stage, and information about them should also be collected when conducting MRCTs. Here, ethnic factors include both Intrinsic and extrinsic factors, and they are well described in ICH-E5 (Ethnic Factors in the Acceptability of Foreign Clinical Data). Briefly, intrinsic factors are the subject's characteristics represented within themselves such as age, gender, race, gene, height, weight. Extrinsic factors are the subject's environment and culture (something outside themselves) which could influence the subject's behavior, practice, and preferences such as tobacco/alcohol use, diet, socio-economic status, and medical practice/standard of care. Based on the understanding of accumulated knowledge about these intrinsic and extrinsic factors, MRCTs should be designed to provide information to support an evaluation of whether the overall treatment effect applies to subjects from participating regions. It is worth noting that such factors are often the reason that regional differences appear. Regions are often a surrogate for information that is unknown. As such, the more that is planned regarding intrinsic and extrinsic factors at the design stage, the more focused examinations of study heterogeneity can be, with less post hoc data dredging to find reasons for a regional difference finding.

Towards the concept of de-emphasizing geographic region in favor of more relevant factors, ICH-E17 (2016) introduces the concept of subpopulation, which is described as a subset of subjects across the regions who are thought to be similar with respect to intrinsic, and/or extrinsic factors relevant to the disease area and/or drug under study. In order to further evaluate consistency of treatment effects, a pooled subpopulation whose members share one or more intrinsic or extrinsic factors deemed important for the drug development program may be particularly useful when regulators (looking to leverage additional data from beyond their local population) would like additional data to be available from a relevant subpopulation to allow generalizability to a specific population within their regulatory country or region. MRCTs conducted according to ICH-E17 (2016) will enable investigation of treatment effects in overall populations with (1) multiple ethnic/intrinsic factors, (2) extrinsic factors, as intrinsic/extrinsic factors described in the ICH-E5 guideline, and (3) geographic region hence enable investigating consistency in treatment effects across populations.

In this chapter, we will pay particular attention to the subpopulation concept introduced in ICH-E17 (2016) and probe the deeper questions such as 'Do we analyze subpopulation differently?', 'Are questions different for 'subpopulation' from 'pooled region' and 'subgroup'?' Then we will review the examples of subpopulation, and finally offer suggestions for the best practice in defining, analyzing, and interpreting subpopulation.

11.2 Definition of Subpopulation

In ICH-E17 (2016), a subpopulation is defined as pooling a subset of the subjects from a particular region with similarly defined subsets from other regions to form a pooled subpopulation whose members share one or more intrinsic or extrinsic factors important for the drug development program. It goes on to say this approach may allow generalizability to a specific population within a regulatory agency's jurisdiction, and hence provide a stronger basis for regulatory decision-making, and the pooled subpopulations may provide a basis for regulatory decision-making for relevant regulatory authorities.

Subpopulation should account for factors within or across geographic regions that may affect the response to the treatment. Such potential intrinsic and extrinsic factors are race, gene, local clinical practice, locally available concomitant medications, and culture. Documenting the definition of subpopulation and the plan for assessment of subpopulation effects at the time of study design also provides appropriate perspectives for both anticipated and any unanticipated regional findings at the study conclusion. Resulting subpopulation accounting for such intrinsic and extrinsic factors together with early data and scientific understanding should more likely be scientifically justifiable and less heterogeneous.

Subpopulation should be defined with the consideration of the outcome of the study. If the objective is to study the pharmacokinetics of the drug (i.e. PK study), intrinsic factors such as race and genetics may be more important than the extrinsic factors to understand how the drug is processed in the subject's body. If the objective is to evaluate efficacy and safety of the drug, extrinsic factors may be more important since local clinical guideline and medical practice, and culture may have more impact on the subject's response to the certain drug of investigation. In fact, a drug's sensitivity to intrinsic and extrinsic factors are well described in ICH-E5 including properties such as linear pharmacokinetics, flat dose-response, a wide therapeutic dose range among others.

ICH-E17 (2016) also mentions that subpopulations should be thoroughly defined and evaluated at the protocol design stage and how this prespecification should help for the relevant regulatory authorities.

11.3 Example

An example of subpopulation using an intrinsic factor is the phenotype of CYP2D6 enzyme that is an important determinant of treatment response for a particular drug. CYP2D6 is responsible for the metabolism and elimination of approximately 25% of clinically used drugs, and if the phenotype is identified as poor metabolizer opposed to an extensive metabolizer, the drug will stay in the body system longer, and it may affect the response to that drug treatment. Therefore, CYP2D6 may be an important intrinsic factor to consider subpopulation for a certain drug.

Another example of a subpopulation using as intrinsic factor, is related to diagnosed ADHD subtype (inattention, hyperactivity, combined) which is an important factor to understand the magnitude of responses to treatment in evaluating the ADHD medications. Patients diagnosed with ADHD inattentive subtype typically exhibits notably smaller response relative to the hyperactive subtype as hyperactivity behaviors are easier to detect (Tanaka et al. 2013). In defining subpopulation, consideration may need to be given to the subject's ADHD subtype across geographic regions.

A well-known example where medical practice played a significant role in resulting regional differences was from the PLATO trial (Wallentin et al. 2009) in evaluating ticagrelor versus clopidogrel in patients with acute coronary syndromes. The study showed an overall statistically significant efficacy effect. However, results in North America and, in particular, in the United States showed a trend in the opposite direction. One potential confounder based on post hoc analysis was the dose of maintenance aspirin that was used at much higher dosage in the US than in the rest of the world. Analyses controlling for aspirin dose (low, medium, high) revealed that what was initially discovered as a region effect could be attributed to differences in aspirin doses being confounded with region. Having considered aspirin dose in the design of the study may have avoided much work to try to explain the regional finding. Indeed, in the follow-up PEGASUS trial, ticagrelor effect with a low-dose aspirin was studied to confirm the efficacy to reduce the risk of cardiovascular death, myocardial infarction, or stroke (Bonaca et al. 2015). Furthermore, no apparent heterogeneity in the efficacy was observed across geographic region including North America (Fig S2 Supplementary Appendix, Bonaca et al. 2015).

ICH-E17 (2016) also recommends that when applicable, PK investigations should be undertaken in subjects from major subpopulations that are intended to be included in MRCTs. For example, Hispanics living in North and South America, or Caucasians living in Europe and North America as an ICH-E17 (2016) subpopulation example. This seems appropriate for a PK study since the objective is to study such an intrinsic factor, in this case, race, to study PK characteristics of the drug. Race is inherent to the subject, and it is not expected to be impacted by external factors. However, for evaluating efficacy and safety of the drug, caution is needed for the extrinsic factors to decide subpopulation since response to the drug may be impacted by external factors. For example, medical practice such as concomitant medications can play a significant role in heterogeneity of treatment outcomes as in the PLATO trial. Another example may be a cultural/societal difference that could impact a subject's response to treatment. For example, a culture may be stoic nature limiting the expression of pain compared to another non-stoic culture. Such a cultural difference may impact the results of a trial that requires measurement of subjective pain scores. Apparent regional differences may actually be cultural differences. An understanding of such cultural differences would be very useful at the design stage of the study as well as in designing the trial to account for such differences (e.g. through a pre-defined subpopulation or through stratification, etc.).

An important parallel to the idea of a subpopulation comes from the book by the National Research Council (2011) on precision medicine. It notes that "Precision Medicine refers to the tailoring of medical treatment to the individual characteristics

of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment". Note the word "subpopulation" used by the NRC, in an analogous fashion to the idea put forth in the draft ICH E17 guidance (ICH E17 2016).

When considered carefully, intrinsic and extrinsic factors can provide insight for estimating the treatment outcome among the different groups that are identified by those factors. Identifying these factors in the planning stage can help to anticipate, plan for, and provide proper perspectives in trial results across regions. Based on accumulated information about intrinsic and/or extrinsic factors and the use of pooled subpopulations may provide useful ways to study different responses to the drug treatment in the context of its intrinsic and extrinsic factors. In fact, subpopulations may be more informative than geographic regions when they identify factors that will be confounded with region. In addition to pooling for subpopulations, ICH-E17 draft guidance (ICH E17 2016) discusses the concept of pooled regions, discussed more in the next section.

11.4 Contrast from Pooled Regions, and Subgroups

ICH-E17 (2016) uses other terms similar to subpopulation, namely, pooled regions and subgroup. Pooled regions are defined as a subset of subjects where data can be pooled together within and/or across geographical regions, countries or regulatory regions based on a commonality of intrinsic and/or extrinsic factors for purpose of regulatory decision-making. Here, the focus appears more for the regulatory review and approvability in the region where the regulatory authority is responsible.

US FDA is known to conduct their own analyses using the trial data that the sponsor submitted, and one of the typical US FDA analyses is to evaluate by-region data such as US vs outside US. In fact, US FDA is not alone, as other regulatory agencies often request by-region data, also dichotomizing the data into two regions, one that is the region where the authority is responsible for and the other being outside of that region. While this evaluation certainly is relevant for the region to understand and describe the efficacy and safety of the drug for the people in the region, it becomes challenging or problematic if the region did not participate in the clinical trial or the sample size in that region is very small. From the statistical perspective, this sets up a series of "us versus the rest of the world" one degree of freedom contrasts, each from a different regulatory agency perspective. Such multiple testing will invite false positives, meanings finding of regional inconsistency where such inconsistency does not actually exist.

To address this case, instead of repeating the clinical trial to include the subjects from the concerned region, if the regions in the completed trial data can be pooled to provide more sample size, or if the characteristics of pooled regions defined by intrinsic and extrinsic factors that are considered similar to the concerned region,

the data from the pooled regions can be a basis for the regulatory review for the concerned region (whether the concerned region did not participate in the trial or the sample size is not sufficient). Pooled data from East Asia, for example, Japan and Taiwan, can possibly be submitted to other regulatory agency in East Asia such as Korea when the subject's response to the drug is considered similar within the subset of subjects, in this case East Asia. It is also common for North America to be pooled region formed from Canada and the United States.

From this perspective, the pooled regions and the subpopulation are indeed distinctive as the pooled regions are driven by the authority's jurisdiction (and often geography) where the subpopulation is more driven by the intrinsic and extrinsic factors that potentially impacts the subject's response to the drug, with the added advantage that the subpopulation will cross common geographical boundaries.

The concept of subgroup is the more familiar terminology than pooled region or subpopulation, as a subgroup analysis to assess consistency of the primary trial result performed among different subset of subjects defined by patient demographics (ex. age, gender) or baseline clinical characteristics (ex. different severity). ICH-E17 (2016) describes that 'of most interest are subgroups defined according to intrinsic and extrinsic factors likely to be prognostic for the course of the disease or plausibly predictive of differential response to treatment'. Examples in ICH-E17 (2016) include subgroups defined by ethnicity (e.g., Asian, Black or Caucasian), medical practice/therapeutic approach (e.g., different doses used in clinical practice) or genetic factors (e.g., polymorphisms of drug metabolizing enzymes).

Guidance is also available from the EMA *Guideline on the investigation of subgroups in confirmatory clinical trials* (EMA 2014): "In recent years the experience has grown that country (or region) can be similar important prognostic factors covering important intrinsic and extrinsic factors, including different attitudes to diagnosis, co-medication and other aspects of the concomitant setting. Although it is recommended to address these aspects by directly addressing the respective variables, country (or region) as an entity for checking the context-sensitivity (or robustness) of the treatment effect is of importance to regional drug licensing bodies and as a plausible source for learning about the robustness of the treatment effect." The guidance in Sect. 5.3 goes on to discuss the importance of intrinsic and extrinsic factors, aligning with the E17 (ICH E17 2016) concept of creating subpopulations. Further, the EMA guideline goes on to describe three scenarios for examining subgroups. Scenario 1 (Sect. 6.3 of the guideline and Annex (1)) describes the scenario where the clinical data presented are overall statistically persuasive with therapeutic efficacy demonstrated globally. It is of interest to verify that the conclusions of therapeutic efficacy and safety apply consistently across subgroups of the clinical trial population. Prespecified subpopulations fall into this category. Scenario 2 (Sect. 6.4 of the guidance and Annex (2)) addresses the scenario where the clinical data presented are overall statistically persuasive but with therapeutic efficacy or benefit/risk which is borderline or unconvincing and it is of interest to identify a subgroup that has not been pre-specified as part of the confirmatory testing strategy, where efficacy and risk-benefit would be convincing. Post hoc subpopulations grounded in good scientific rationale (external to the clinical trial including successful results in other

trials) with a biologically plausible explanation and a large benefit could fall into this scenario. Scenario 3 addresses the situation where the clinical data presented fail to establish statistically persuasive evidence but there is interest in identifying a subgroup, where a relevant treatment effect is evident and there is compelling evidence of a favorable risk-benefit. Scenario 3 could be a situation where a subpopulation is established to be further tested in future trials.

There may be an overlap when trying to distinguish between subgroup and subpopulation. However, it appears that the subpopulation is a specific type of subgroup that is a pooled subset of the subjects across regions sharing one or more intrinsic or extrinsic factors that may be associated with differential treatment response. Indeed, ICH-E17 (2016) describes that the pre-specified subgroup analyses for study subpopulations that are defined beyond geographical boundaries and based on common intrinsic and/or extrinsic factors may be useful for generating key scientific evidence to support a regional marketing authorization. From the PLATO study example, it was discovered and labeled in the US that the dose of maintenance aspirin (high, low) was associated with differential treatment response. This aspirin dose was explored when the by-region analysis resulted in qualitative interaction indicating the treatment effect in North America was much smaller than other regions. In this example, both geographic region and aspirin dose were subgroups, but the aspirin dose perhaps could have been considered more as a subpopulation since the determination of aspirin dose is related to local clinical practice in different regions, can be recognized at the design stage, can be pooled across regions, and the effect on the treatment response has biological plausibility.

In summary, subgroup analyses should be pre-specified with the type of subgroup including pooled regions and pooled subpopulation. In identifying a pooled subpopulation, cautious scientific investigation is needed to evaluate a list of relevant intrinsic and extrinsic factors to delineate the ones that have the potential to influence a subject's response to the concerned drug. Pooled regions are more related to geographical and political boundaries of the concerned regulatory authority.

If subgroup differences are observed unexpectedly, then a post hoc examination of subgroup differences across regions (or pooled regions) is warranted. To help give integrity to such post hoc examinations, things to consider are the biological plausibility, internal/external consistency (different endpoints, different trials, same drug class), strength of evidence (marginal or overwhelming), and the statistical uncertainty (Possibly random chance?).

11.5 Best Practice

The goal of ICH-E17 (2016) is to minimize the occurrence of unexpected results driven by regional heterogeneity, hence to increase the acceptability of the trial data in global regulatory submissions. While the current public draft of the ICH E17 guidance (ICH E17 2016) does not give specific methodology, the EMA subgroup guideline (EMA 2014) offers various scenarios for subgroup analyses, and the con-

cept of subpopulation to strengthen understanding of a trial's results falls within scenario 1 and 2 of the subgroup guideline. Disparate, inconsistent regional findings in a trial that is overall successful makes interpretation of the trial difficult for health authorities in the region. Throughout the ICH-E17 draft guidance (ICH E17 2016), emphasis was given to intrinsic and extrinsic factors in planning and designing the multi-regional clinical trials. In particular, the concept of a subpopulation was introduced in defining a particular subgroup by pooling a subset of the subjects across regions who share one or more intrinsic and extrinsic factors that potentially influence a subject's response to the drug.

To operationalize this concept, hierarchical layout may enable to sort out the steps to form the subpopulation and analysis plan. First, intrinsic and extrinsic factors can be listed as a checklist that enables the structured assessment of factors to identify the ones that potentially contribute to heterogeneity in the subject's treatment response. While the study protocol standardizes the study population and conduct of the study, the sensitivity of subject's response to the treatment still needs to be examined in multi-regional clinical trials by a thorough consideration of intrinsic and extrinsic factors.

The second step is to examine the geographic regions expected to participate in the trial to form the pooled regions. Since the pooled regions are more related to geographical and political boundaries of the concerned regulatory authority to evaluate the approvability of the drug, consideration needs to be given to the region that is not participating the trial to see if the pooled regions can provide the sufficient data that can be the basis for approvability. The illustrated example was to use the pooled data from Japan and Taiwan, for the regulatory review for Korea (which did not participate in the trial) when the subject's response to the drug is considered similar within the subset of subjects in East Asia.

The third step is to form the subgroup, which consists of intrinsic, and extrinsic factors or geographic regions that are not accounted for in the subpopulation or pooled regions in the previous two steps. Ideally prespecification of the subgroup is preferred, but the findings from the trial may necessitate the ad hoc subgroup to be formed.

Once the subpopulation, the pooled regions, and the subgroup are identified, the next step is to take preventative actions, for example, by modifying the study design to accommodate particular factors for certain country/region or a priori plan analyses using these factors. In addition, this can be useful to target country-specific training or, in extreme cases, while at the design stage to choose to exclude a country or region from the trial.

It should be noted that the E17 guidance (ICH E17 2016) focuses on the design of MRCTs; the EMA subgroup guidance (EMA 2014) offers specific guideline on the subgroup analyses in several different scenarios. These include the analyses of confirmatory subpopulation, regional consistency, supportive subgroup and exploratory subgroup evaluations.

References

Bonaca, M., Bhatt, D., Cohen, M., Steg, P., Storey, R., Jensen, E., Magnani, G., Bansilal, S., Fish, M., Im, K., Bestsson, O., Ophuis, T., Budaj, A., Theroux, P., Ruda, M., Hamm, C., Goto, S., Spinar, J., Nicolau, J., Braunwald, E., Sabatine, M., & For the PEGASUS-TIMI 54 sterring committee and investigators. (2015). Long-term use of ticagrelor in patients with prior myocardial infarction. *New England Journal of Medicine, 372*, 1792–1800.

European Medicines Agency (EMA). (2014). *Guideline on the investigation of subgroup in confirmatory clinical trials (draft)*. London. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf.

International Conference on Harmonisation (ICH). (1988). *E5 guideline on ethnic factors in the acceptability of foreign clinical data*. Geneva.

International Conference on Harmonisation (ICH). (2006). *E5 (R1) implementation working group, questions & answers*.

International Conference on Harmonisation (ICH). (2016). *E17 draft guideline on general principles for planning and design of multi-regional clinical trials.* Geneva.

International Conference on Harmonisation (ICH). (2014, July). *E17 final concept paper*. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E17/E17_Final_Concept_Paper_July_2014.pdf.

National Research Council (NRC). (2011). *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*. http://nap.edu/13284.

Tanaka, Y., Rohde, L. A., Jin, L., Feldman, P. D., & Upadhyaya, H. P. (2013). A meta-analysis of the consistency of atomoxetine treatment effects in pediatric patients with attention-deficit/hyperactivity disorder from 15 clinical trials across four geographic regions. *J Child Adolesc Psychopharmacol, 23,* 262–270.

Wallentin, L., Becker, R. C., Budaj, A., et al. (2009). Ticagrelor versus clopidogrel in patients with acute coronary syndromes. *New England Journal of Medicine, 361,* 1045–1057.

Chapter 12
Adaptive Group-Sequential Multi-regional Outcome Studies in Vaccines

Inna Perevozskaya

12.1 Background: Case Study

This work was motivated by the need to design and de-risk a very large study evaluating safety and efficacy of an experimental vaccine designed to prevent post-operative invasive infections with *Staphylococcus aureus* (*S aureus*). *S aureus* is a leading cause of healthcare-associated infections, resulting in a substantial burden to health care systems. It is a particularly challenging pathogen, resistant to antibiotics and capable of causing a wide spectrum of symptoms, ranging from mild skin infections to deep wound and surgical site infections, bacteremia and sepsis, potentially leading to death (Pfizer, 2015).

In the US, about 20% of all post-surgical infections at the incision site are attributed to *S aureus.* Patients who suffer such infections due to antibiotic resistant (MRSA) or antibiotic sensitive (MSSA) *S aureus* have worse clinical outcomes, including increased mortality in comparison with non-infected patients (Pfizer, 2015).

At the time this study was designed, there were no licensed vaccines to prevent *S aureus* infection, so it was a serious and unmet medical need. Therefore, the vaccine under investigation was granted Fast Track designation by the U.S. Food and Drug Administration (FDA) in February 2014 (Pfizer, 2015). In such situations, a submission based on one pivotal study (rather than two typically required) may be sufficient to get approval if the evidence is compelling. The competitive landscape also made time-to-market a crucial factor in the development. All of the above contributed to pressure on the study team to accelerate clinical development of this vaccine candidate using innovative techniques, to use resources wisely and manage risks appropriately. This presented a prime opportunity to explore adaptive design strategies, both at the study and at the program levels.

I. Perevozskaya (✉)
Statistical Innovation Group, GSK, 1250 S Collegeville Rd, Collegeville, PA 19426, India
e-mail: inna.x.perevozskaya@gsk.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_12

The primary outcome of this study was defined as postoperative *S aureus* blood stream infections and/or deep incisional or organ/space surgical site infections occurring within 90 days after elective posterior instrumented lumbar spinal fusion (ClinicalTrials.gov: https://clinicaltrials.gov/show/NCT02388). About 10–60 days prior to their scheduled surgery, patients would be randomized (in a 1:1 ratio) to either the *A aureus* vaccine (also labeled SA4Ag) or placebo. The primary objective of the study was demonstrating that the number of patients developing above-mentioned type of infections is lower in the vaccine group compared to the placebo group.

12.2 Overview of Statistical Approaches to Vaccine Efficacy Studies

In a randomized placebo-controlled trial where the endpoint of interest is occurrence of infection, vaccine efficacy (VE) is defined in terms of relative risk (RR):

$$VE = 1 - RR \tag{12.1}$$

There are three classes of commonly used methods for design and analysis of vaccine efficacy studies (Nauta, 2010); they differ in how relative risk is defined and how a subject's follow-up is accounted for. We will describe them briefly to allow understanding of the subsequent material. For more statistical details, we refer the reader to Chap. 7 of Nauta (2010).

1. ***RR* as a ratio of 2 attack rates ("binomial model")**: The attack rate is the risk of an infection-free subject getting infected during a fixed follow-up period. This method assumes all subjects are followed up for a fixed pre-defined amount of time. So the incidence of infection can be simply characterized by presence or absence (binary endpoint) of the event of interest at the end of the pre-defined follow up time period. The attack rate π is defined as ratio of number of subjects developing infection (over pre-specified follow-up period) to the total number of subjects exposed and initially infection-free. Let π_{trt} and π_{pbo} represent attack rates in the vaccine and placebo treatment groups, respectively. Under this setting $RR = \theta = \pi_{trt}/\pi_{pbo}$ and both π_{trt} and π_{pbo} can be interpreted as binomial proportions (i.e. incidences of infection in the two treatment groups follow their respective binomial distributions).
2. ***RR* as ratio of two infection rates ("Poisson model"):** The infection rate is the risk of experiencing an infection during a given time unit, such as month or a year. Relative risk *RR* is defined the same way as in the previous model, i.e. $RR = \theta = \pi_{trt}/\pi_{pbo}$, but the rates π_{trt} and π_{pbo} are now rates per person-years, accounting for variable exposure. For example, $\pi_{trt} = s/T$, where s is number of cases in the vaccine group and T is total person-time of all subjects in the vaccine group, which in turn is defined as sum of follow-up times for all subjects in that group (from enrollment until event or termination of the study or a drop-out). This

method of capturing and describing the incidence of infection is more tailored to practical applications when subject's time of entry into the study and length of follow-up vary. It is usually assumed that number of cases in each treatment group follows a Poisson distribution. The cornerstone assumption underlying this model is that the risk of infection is constant over the time of follow-up. This, in turn requires that the event rate is low and the infectious disease does not have seasonality. It's important to note that under fixed follow-up scenario (and low event rate) this model becomes very similar to the binomial model described earlier. In other words, if the infection rate is low, it will *approximately* equal to the attack rate. We will use that assumption in our trial design to facilitate the interpretation of vaccine efficacy estimates.

3. ***RR* is ratio of two forces of infection rates ("time-to-event model")**. This approach is based on classical survival analysis, where survival function $S(t)$ is modeling time from enrollment to infection. Force of infection is equivalent to hazard function, which is typically used in survival analysis to compare two treatment groups. The relative risk *RR* is defined as a ratio of two force of infection rates between the two treatment groups: $RR = h_{trt}(t)/h_{pbo}(t)$. This approach is the most complicated of all; it is recommended for use when constant attack rate over time cannot be assumed.

In this study, the background occurrence of staph infections post-surgery without vaccine (i.e. placebo attack rate) was expected to be fairly low: about 2–3% in the selected patient population. Also, fixed surveillance period (90 days post-surgery) was employed, making the second approach utilizing Poisson assumptions suitable for this study design. Specifically, if we express vaccine efficacy using percentage scale (0–100%), as often done by clinicians, rather than probability scale (0,1), we can formulate the problem the following way:

$$VE = (1 - \theta) * 100\% = \left(1 - \frac{\pi_1}{\pi_0}\right) * 100\% \tag{12.2}$$

$$\pi_1 = \frac{s_1}{T_1},\ \pi_0 = \frac{s_0}{T_0} \tag{12.3}$$

where s_0, s_1 are number of events (i.e. cases of staph infection) in the placebo and vaccine treatment groups, respectively, and T_0, T_1 are exposures expressed in person-years. Under this setting, the number of cases in each treatment group follows a Poisson distribution:

$$s_1 \sim Poi(\lambda_1),\ s_0 \sim Poi(\lambda_0) \tag{12.4}$$

Both rates λ_0 and λ_1 are unknown, but the question of interest is to estimate only the ratio λ_1/λ_0, making λ_0 a nuisance parameter. To get around that, it is a standard practice in vaccine trials to use exact binomial inference based on conditional distribution for trial planning and inference: conditioning on the total number of events

$s = s_0 + s_1$, the number of events coming from the vaccine group follows a binomial distribution Nauta (2010) with parameter π:

$$s_1 \sim \text{Bin}(s, \pi), \quad \pi = \frac{T_1\lambda_1}{T_1\lambda_1 + T_0\lambda_0} \tag{12.5}$$

If we recall that our parameter of interest is relative risk $\theta = \lambda_1/\lambda_0$, denote $r = T_1/T_0$ and apply equal randomization ratio in combination with fixed follow up periods (i.e. $T_1 = T_0$), we can re-write (12.5) as:

$$\pi = \frac{T_1\theta}{T_1\theta + T_0} = \frac{r\theta}{r\theta + 1} = \{r = 1\} = \frac{\theta}{\theta + 1} \tag{12.6}$$

or, alternatively:

$$\theta = \frac{\pi}{1-\pi}, \quad VE = \left(1 - \frac{\pi}{1-\pi}\right) * 100\% \tag{12.7}$$

Equations (12.6) and (12.7) are key to understanding the relationship between vaccine efficacy and parameter π. They provide an easy "switch" between two-sample Poisson problem and one-sample binomial problem formulations. The latter is used to power the trial and to perform statistical inference. However, it is not very useful for final results presentation. Since parameter π is hard to interpret clinically, its point estimates and confidence intervals need to be "back-transformed" to the corresponding quantities for VE (via θ) using Eq. (12.7). Such exercise provides results on the same scale as the original research question, i.e. expressed in terms of vaccine efficacy rather than abstract conditional distribution parameter.

Clinicians usually prefer to state study hypotheses in terms of vaccine efficacy. Vaccine trials are unique in the sense that they often carry an additional requirement of "super-efficacy", i.e. it's not enough to demonstrate that vaccine is simply better than control, it needs to be better by a certain amount:

1. $H_0 : VE \leq 0$ versus $H_A : VE > 0$(without "super − efficacy")
2. $H_0 : VE \leq \delta$ versus $H_A : VE > \delta$(e.g. $\delta = 30\%$ for "super − efficacy")

Given relationship between VE and π described in (7), these hypotheses can be re-written as:

1. $H_0 : \pi \geq 0.50$ versus $H_A : \pi < 0.50$(without "super − efficacy")
2. $H_0 : \pi \geq 0.41$ versus $H_A : \pi < 0.41$(e.g. $\delta = 30\%$ for "super − efficacy")

When conditional distribution is used in power calculations, the process becomes an iterative search for the "optimal" pair (s, s_1), satisfying desired Type 1 error and power requirements. Various values of s are examined in increasing order; for each s, an "optimal" cut-off s_1 is determined such that $(1-\alpha) \times 100\%$ lower confidence

interval bound (LCB) for *VE* exceeds δ (super-efficacy threshold) and power is at the desired level, e.g. 80%. Power is computed as probability to observe $\leq s_1/s$ events in the vaccine group under the alternative hypothesis H_A, (e.g. assuming $VE = 80\%$). All calculations are carried out using *Bin(s,* π*)* distribution described in (12.5). In this process, the power is driven by total number of events *s* accrued. Smaller values of *s* may be too crude to achieve required Type 1/Type 2 error levels; the search continues by increasing *s* until such conditions can be met. It's important to note that the placebo background event rate λ_0 does not impact these power calculations at all. By conditioning on total number of events *s*, we have eliminated the need to worry about nuisance parameter placebo event rate λ_0 in our power calculations. However, the trade-off is that the study sample size of *N* patients now becomes a random variable following a Poisson distribution, which depends on λ_0, i.e. its dependency on a nuisance parameter cannot be eliminated. This phenomenon is similar to survival studies where the information is driven by number of events accumulated rather than number of patients. Similar to survival studies, the placebo event rate can be very impactful on the actual sample size *N*. In our case the expected number of patients *N* required to accrue *s* events would depend on *VE* and λ_0, in addition to *s*:

$$E(N) = 2 * \frac{s}{\lambda_0 * (2 - VE)} \tag{12.8}$$

If λ_0 is small, the sample size *N* becomes very large; that's why "rare" diseases like the one in our study pose a challenge. On the other hand, if vaccine is very effective, i.e. *VE* close to 100%, the sample size increases as well because it will be harder to accrue required *s* events because the vaccine will prevent them from occurring in the vaccine group. Both *VE* and λ_0 are unknown at the time of study design and both can impact the sample size *N* quite dramatically. Usually there is a great deal of uncertainty about them. In this study, the plausible values of *VE* were 50–80% with $VE = 60\%$ being the "target" commercially viable value; the plausible values for λ_0 ranged from 0.06 to 3%, according to the literature.

We will illustrate the impact of relationship (12.8) on actual sample size by computing expected sample sizes for various plausible values of these parameters (Table 12.1).

It is evident from Table 12.1 that in order to demonstrate super-efficacy of SA4Ag vaccine a very large multi-center trial would be required. Both placebo event rate and true vaccine efficacy were highly influential in sample size calculation and unknown at the time of study design.

In addition to that challenge, study population heterogeneity presented a separate challenge of its own:

- Placebo attack rate λ_0 varied considerably across patient sub-populations (surgical subgroups) because it depends on type of surgery performed and other underlying co-morbidities.
- Multiple centers/regions contribute to variability in primary outcome (infection).
- As with any surgical procedure, surgeon's skill level and quality of post-operative care could affect the primary outcome (infection) and contribute to its variability.

Table 12.1 Impact of VE and Placebo attack rate on sample size

($\alpha = 0.05$, power 90%, super-efficacy $\delta = 30\%$)							
Sc#[a]	VE	λ_0	s	s_1	s_0	n_1	n_0
1	0.6	0.03	154	51	103	3667	3667
2	0.6	0.02	154	51	103	5501	5501
3	0.6	0.01	154	51	103	11,001	11,001
4	0.7	0.03	69	20	49	1770	1770
5	0.7	0.02	69	20	49	2654	2654
6	0.7	0.01	69	20	49	5308	5308
7	0.8	0.03	38	9	29	1056	1056
8	0.8	0.02	38	9	29	1584	1584
9	0.8	0.01	38	9	29	3167	3167

[a]Scenario number

Since the proposed study would be the first one to evaluate efficacy in patients where underlying parameters were poorly known and would further require a large sample size, the study team was challenged to "de-risk" development by utilizing an adaptive design.

The initial adaptive design was very complicated and had multiple objectives: it was attempting to address multiple surgical sub-populations in one study as well as uncertainty in treatment effect and placebo attack rate via sample size re-estimation and population enrichment. After many deliberations, it was decided to separate the heterogeneity issue from parameter uncertainty and put the former one aside, focusing on one surgical sub-population only (elective spinal fusion), which was believed to have the highest background event rate according to the literature. This strategy would allow the team to study and confirm vaccine efficacy more quickly and then augment the population in a separate study, should the vaccine efficacy turn out to be promising. Such stage-wise investment was necessary because previous studies of this vaccine provided only immunogenicity data, not actual *VE* in patients, so all assumptions on vaccine efficacy were based on translation of immunogenicity data into perceived *VE* and needed to be confirmed by an actual clinical study data. Even this "fast" study focusing on spinal fusion subpopulation would have been long and expensive, so additional de-risking options, including sample size re-estimation and early futility stop, were examined. We will review the evolution of various adaptive design options in the following section.

12.3 Clinical Development Strategy Options

12.3.1 Adaptive Phase 2/3 Design

In an attempt to accelerate time to market, one initial program development option included skipping Phase 2 study entirely, i.e. proceeding straight to Phase 3 after immunogenicity studies. The objectives typically addressed in Phase 2, such as initial evaluation of vaccine efficacy and getting a firmer estimate on the background infection rate, and objectives typically addressed in Phase 3 (e.g. confirming vaccine efficacy with super efficacy alternative) were combined within a single study similar to a seamless Phase 2/3 design. It was proposed to address these objectives via multiple interim analyses with options for unblinded sample size re-estimation and early futility stops.

The study was designed assuming $VE = 60\%$, super-efficacy threshold $\delta = 30\%$, desired power of 90% and type 1 error of $\alpha = 0.05$ (two-sided). The underlying placebo event rate was assumed to be $\lambda_0 = 0.02$. Following the sample size calculation procedure based on conditional distribution described in Sect. 12.2, a fixed trial to establish super-efficacy of the vaccine over placebo at a given super-efficacy threshold δ would require approximately $N = 5501$ patients enrolled per treatment arm to accrue $s = 154$ total events required to achieve the desired power. The high sample size was primarily driven by low placebo rate and super-efficacy threshold of 30%. At the end of the study, if at most 51 cases of infection (out of total $s = 154$) were observed in the vaccine treatment arm, the null hypothesis could be rejected and the study would be declared a success. The cut-off point of $\leq$51/154 for the observed proportion is equivalent to $(1-\alpha) \times 100\%$ LCB on VE to be above 30%, using relationship between VE and π in Eq. (12.7).

To de-risk this large study and fill-in the placeholder of a "skipped" phase 2 study, an early interim analysis (IA) for futility based on conditional power was proposed after $s = 16$ total events would be observed (~10% of total information planned for this study).

Conditional power is a widely used concept to make interim decisions in an adaptive clinical trial. It quantifies the probability of achieving final study success criterion, given observed interim data and initial assumptions about the treatment effect. The basic principle of futility assessment is early study termination if conditional power is low (<30%), proceed without modification it is high (>80%) and possibly increase sample size after interim analysis (IA) if the probability of final success is modest (30–80%), given the current (interim) data. The team has considered this and other types of adaptation (such as population enrichment) at later time points in the study (IA $\geq$ 2). At the time of the 1st IA only early futility stop would be considered.

For simplicity, let's assume this first futility look is the only interim look in the study. It is common to refer to "event split" in vaccine efficacy trials based on total event count. That is, if the total number of events is $s = 16$, the "split" is characterized by a pair (s_1, s_0) where s_1 and s_0 describe number of those events occurring in the

vaccine and placebo groups, respectively. Since $s_0 + s_1 = 16$, its sufficient to track s_1 (vaccine cases) only. In other words, given total number of cases, the "split" tells the story of how effective the vaccine is: the lower is s_1, the more effective the vaccine is. The final success criteria for this study was defined as observing $\leq$51/154 events at the end of the trial. So, using the notation above the conditional power (CP) can be defined 2 ways:

$$\text{CP1} = \Pr(\text{to see} \leq (51 - \mathbf{s_1})/(154 - \mathbf{16}) \text{ events in vaccine group} \,|\, \text{VE} = 0.6) \tag{12.9}$$

$$\text{CP2} = \Pr(\text{to see} \leq (51 - \mathbf{s_1})/(154 - \mathbf{16}) \text{ events in vaccine group} \,|\, \text{VE estimated}) \tag{12.10}$$

Both CP expressions depend on assumptions and observed interim data but to a different degree: CP1 is more dominated by "belief" that VE is consistent with what was hypothesized, i.e. VE = 0.6 in our case, which was the alternative hypothesis used to power the study. The CP2 rule is more sensitive to the interim data and its departures from assumed values because that data enters the equation twice: first in the final success criteria part highlighted **in bold** in both equations and second as a part of estimated *VE* used to calculate the probability of "success". The futility stop is triggered if CP falls below 30%. The two CP rules can "induce" 2 different designs:

- Design 1: if CP1<30% after 16 events then stop for futility
- Design 2: if CP2 < 30% after 16 events then stop for futility

A sample illustration of what kind of "split" (among 16 cases) would be required to trigger a futility stop under each of these designs is given in Table 12.2. We can quickly see that CP1 rule which "believes" in VE = 60% would require quite an extreme split going into direction of vaccine harm [e.g. (15,1)] to declare futility, while the CP2 rule, which places more emphasis on the interim data, would stop for futility at much less extreme split of (6,10). The latter split corresponds to an estimated VE = 40%. The reason CP2 rule appears more aggressive is because it incorporates interim data suggesting efficacy is not much more than 30%; since the final VE estimate must be significantly greater than 30%, a VE point estimate below 40% means the super-efficacy goal is unlikely to be achieved.

A simple table like this one illustrates how decision rules based on CP translate into subsequent estimates of VE observed, helping the team to make some sense of the rules they were proposing. It became quite obvious that a decision rule based on 10% of information can be a very crude one (either not stopping at all when needed or stopping too aggressively, depending on which CP rule we use). The reason for this lies in highly variable interim estimates of treatment effect at the early information time point. To see how stability of interim estimates can be improved, the team considered IA at later time points such as $s = 24$ and $s = 40$. Following the same procedure as for $s = 16$, the resulting interim decision rules for these designs would have been as follows:

Table 12.2 Illustration of CP1 and CP2 decision rules for total S = 16 cases and various split scenarios

Total	s1-(vac)	s0-(pbo)	ve-alt	pi-alt	CP1	pi est	ve est	CP2
16	0	16	0.6	0.29	0.99	0.00	1.0	1.00
16	1	15	0.6	0.29	0.98	0.06	0.9	1.00
16	2	14	0.6	0.29	0.97	0.13	0.9	1.00
16	3	13	0.6	0.29	0.95	0.19	0.8	1.00
16	4	12	0.6	0.29	0.93	0.25	0.7	0.99
16	5	11	0.6	0.29	0.91	0.31	0.5	0.73
16	6	10	0.6	0.29	0.87	0.38	0.4	0.14
16	7	9	0.6	0.29	0.83	0.44	0.2	0.0
16	8	8	0.6	0.29	0.78	0.50	0.0	0.0
16	9	7	0.6	0.29	0.72	0.56	-0.3	0.0
16	10	6	0.6	0.29	0.66	0.63	-0.7	0.0
16	11	5	0.6	0.29	0.59	0.69	-1.2	0.0
16	12	4	0.6	0.29	0.51	0.75	-2.0	0.0
16	13	3	0.6	0.29	0.44	0.81	-3.3	0.0
16	14	2	0.6	0.29	0.36	0.88	-6.0	0.0
16	15	1	0.6	0.29	0.29	0.94	-14.0	0.0
16	16	0	0.6	0.29	0.23	1.00		0.0

If IA is conducted after s = 24 cases:

- CP1 rule: terminate if ≥18/24 in vaccine group
- CP2 rule: terminate if ≥9/24 in vaccine group

If IA is conducted after s = 40 cases:

- CP1 rule: terminate if ≥22/40 in vaccine group
- CP2 rule: terminate if ≥15/40 in vaccine group

The resulting operating characteristics of these decision rules are summarized in Table 12.3.

This exercise of quantifying operating characteristics of 2 possible designs and varying the time of interim analysis helped to convince the team that 10% of information (16 total events) was too early for an interim review. Futility criteria for such designs did not perform well across the range of vaccine efficacy scenarios: it stopped too infrequently for a non-efficacious vaccine in case of CP1 (VE ≤30% is all part of the "null hypothesis" parameter space here) or terminating efficacious vaccine too easily in case of CP2 (VE ≥30% is desired efficacy range). The only way to get a reasonable performance with these kind of rules was to increase the timing of 1st IA

Table 12.3 Operating characteristics of 2 CP rules, by VE scenario and IA timing

Probability of interim stop under various scenarios of IA1 timing and true VE							
Futility interim timing (total events)	True VE = 0	True VE = 0.1	True VE = 0.2	True VE = 0.3	True VE = 0.4	True VE = 0.5	True VE = 0.6
Using CP1 rule (*based on effect size in alternative hypothesis*)							
S = 16	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S = 24	0.01	0.01	0.00	0.00	0.00	0.00	0.00
S = 40 (24% info)	0.32	0.21	0.12	0.05	0.02	0.00	0.00
Using CP2 rule (*based on effect size in alternative hypothesis*)							
S = 16	0.90	0.85	0.79	0.70	0.59	0.46	0.29
S = 24	0.93	0.87	0.81	0.71	0.58	0.41	0.22
S = 40	0.96	0.92	0.85	0.74	0.57	0.34	0.14
S = 60	0.98	0.97	0.91	0.79	0.60	0.34	0.11

to 40% of information, i.e. at least s = 60 total cases. That finally convinced the team that a proper phase 2 study separate from a phase 3 may be needed. The seamless Phase 2/3 option with early futility look at $s = 16$ events was rejected due to poor performance and the team moved on to designing the next option with a separate Phase 2 study.

12.3.2 Simple Group-Sequential Phase 2b Design Followed by Pivotal Phase 3 Study

After the seamless Phase 2/3 option was rejected, the team focused on a single objective of demonstration a "proof-of-principle" with respect to vaccine efficacy while acknowledging that robust evaluation of vaccine efficacy would be performed later (conditional on proof-of principle success) in a separate pivotal Phase 3 study. More complex adaptations such as unblinded sample size reassessment would be saved for that Phase 3 study while the current Phase 2 study would utilize simpler group-sequential design methodology of Jennison and Turnbull (2000).

The new study was designed using less rigid criteria (as it would be customary for any Phase 2):

- No super-efficacy requirement ($\delta = 0$)
- More relaxed power assumptions ($\alpha = 0.10$, 1 sided, power = 80%)
- Same VE = 60% and placebo background rate $\lambda_0 = 0.02$ as in the previous design

These assumptions led to a fixed design sample size of N = 1858 patients (total) in order to obtain $s = 26$ overall events. At the end of the study, if ≤9/26 events were observed in the vaccine group, the null hypothesis could be rejected (i.e. 90% LCB on VE would be above zero and "proof-of-principle" would be established), triggering Phase 3 development. Even this relatively "small" study would require ~2000 patients to establish proof of principle. That's why adding an interim analysis for futility was considered necessary to further de-risk the study.

As with any interim look, final hypothesis testing had to be adjusted to reflect multiplicity arising from an early interim look. We followed the framework described in Jennison and Turnbull (2000) Chap. 12, utilizing classical group-sequential methodology which is considered "well-understood" by draft FDA guidance (FDA, 2010). Initially, the futility stopping boundary was derived using the normal approximation to binomial distribution. In case of small sample size like the one in this design, the actual power and type 1 error may not hold as "promised" by calculated boundary and the latter needs to be refined using exact binomial calculations. The next section will illustrate this with a working example.

12.3.2.1 Derivation of Exact Binomial Stopping Boundaries and Related Operating Characteristics

The starting sample size $s = 26$ (total events) was obtained for a fixed design using exact binomial calculations (iterative procedure described in Sect. 12.2) to guarantee 80% power and $\alpha = 0.10$ type 1 error. We took that design as a starting point and added 1 interim look at 50% of information with early stopping for efficacy or futility. It is illustrated using EAST 6.4 software below but can be done using any other software of choice. A summary of input parameters is given in Fig. 12.1. It is important to remember that the null and alternative hypotheses have to be stated in terms of parameter π when using conditional distribution setting. The sample size of $s = 26$ was kept fixed and power computed as a function of it, which turned out to be slightly different (0.821) than originally specified 0.8. That happened because the calculations in EAST were based on normal approximation. The main purpose of going through this exercise was not to confirm power but rather to obtain stopping boundaries for a group-sequential design with one look. O'Brien-Fleming type of boundaries were used for both efficacy and futility stopping. They are summarized in Figs. 12.1 and 12.2.

From this group-sequential design with one interim look, we can obtain the numerical values of stopping boundaries (on the proportion scale) which are reproduced in Fig. 12.2 and Table 12.4.

Boundaries obtained from EAST on the proportion scale were converted to number of events (using $s = 13$ or $s = 26$ events, respectively). Obviously, the number of events should be integers, so these intermediate values were rounded-off to the nearest integer in the direction of "outside" of the boundary, i.e. more extreme and difficult to stop. For example, at the efficacy boundary at interim look, 2.798 was rounded-off to 2, while for the futility boundary at the same look, 6.270 was rounded

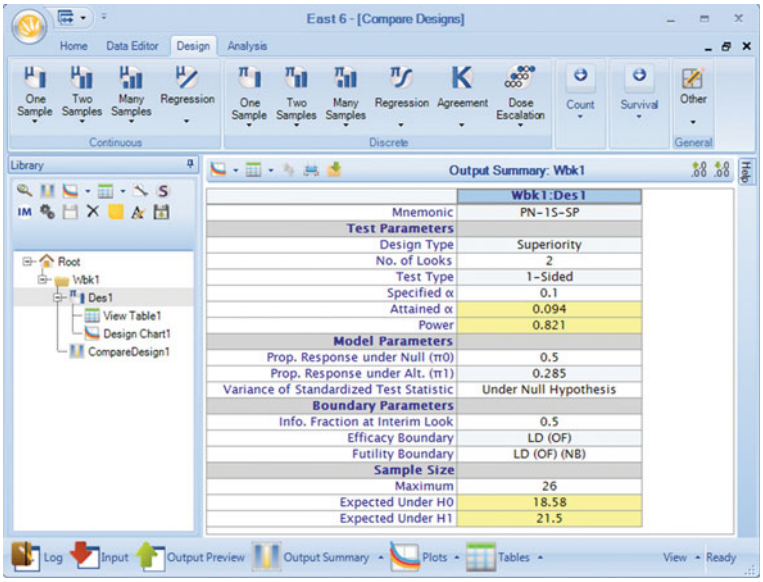


Fig. 12.1 Summary of input parameters in EAST 6.4 for initial group-sequential design (based on normal approximation to binomial distribution)

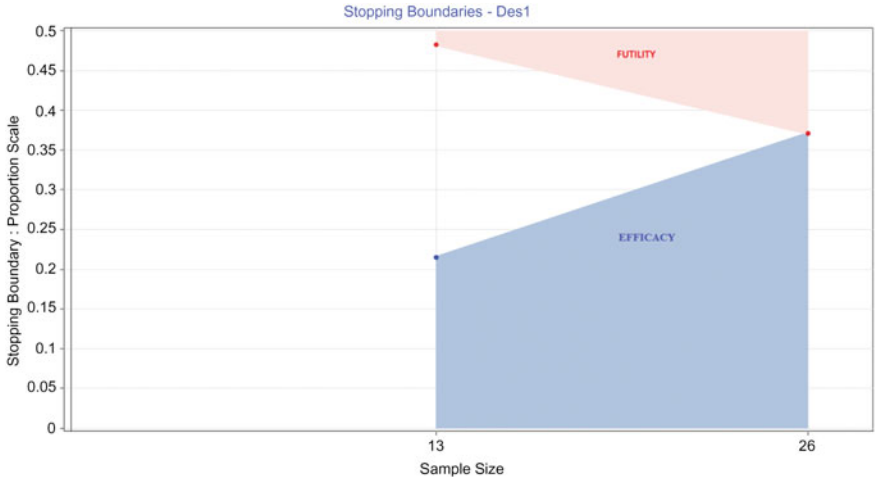


Fig. 12.2 Boundaries of group-sequential design based on normal approximation

off to 7. There are no hard rules about how to round-off: one could have chosen values of 3 or 6 instead and evaluate respective operating characteristics. Our choices were driven by the desire to stop conservatively both for efficacy and for futility. At the final analysis, unlike the continuous case where two boundaries converge to a single

Table 12.4 Boundary values

		Boundary: (prop.)		Boundary (events)		Boundary (GsDesign[a])	
Look #	S. Size	Efficacy	Futility	Efficacy	Futility	Efficacy a (lower)	Futility b (upper)
1	S = 13	0.215	0.482	2.798	6.270	2	7
2 (final)	S = 26	0.371	0.371	9.644	9.644	9	10

[a]See text below for explanation of GsDesign

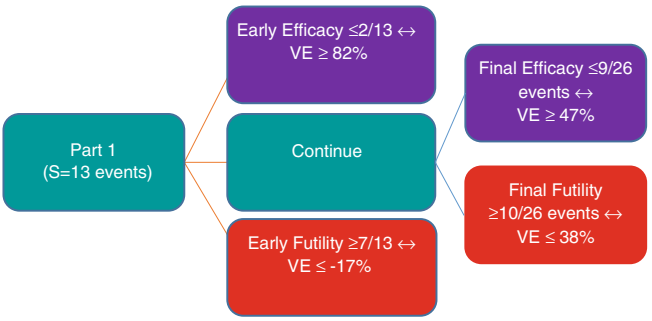


Fig. 12.3 Phase 2 group-sequential design with early stopping for efficacy/futility

point, the discrete boundary values must differ by one, reflecting the discrete nature of the event counts. The discreteness eliminates the "indecision zone", allowing the trial to always come to some conclusion (either success or failure).

The 'discretized' boundaries (last 2 columns of Table 12.3) formed the basis of a new group sequential design, where decision would be based on observed number of events rather than observed proportion of events. Operating characteristics of the former were evaluated using gsDesign R-package (Anderson, 2016) with the following call:

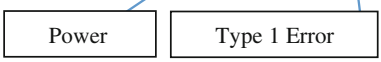
$$Gsbin < -gsBinomialExact(k = 2, theta = c(0.5, 0.285),$$
$$n.I = c(13, 26), a = c(2, 9), b = c(7, 10).$$

The function gsBinomialExact in the gsDesign package calculates probabilities of crossing boundaries a or b ("discretized" boundaries from Table 12.3). The timing of interim analysis and total sample size are specified by vector *n.I*. The vector theta captures null and the alternative hypotheses for the proportion parameter π. The flow-chart of the resulting design with associated decision rules and operating characteristics (computed via above call to gsDesign) are summarized in Fig. 12.3 and Table 12.5, respectively.

In contrast with the original Phase 2/3 proposal described in Sect. 3.1, this new "Phase 2 only" design was very simple: it had only group-sequential early stopping and no sample size re-estimation/population updates. The new study proposal was

Table 12.5 Operating characteristics of Ph2 GSD under variety of VE scenarios

	π	Pr (early eff. Stop)	Pr (early fut. Stop)	Pr (final success)	Pr (final failure)	Overall PR of Success	Ave # Events
VE=0%	0.500	0.011	0.500	0.073	0.415	0.085	19
VE=10%	0.474	0.018	0.424	0.115	0.443	0.133	20
VE=30%	0.412	0.049	0.257	0.267	0.428	0.315	22
VE=50%	0.333	0.139	0.103	0.498	0.260	0.638	23
VE=60%	0.286	0.235	0.049	0.575	0.141	0.810	22
VE=70%	0.231	0.393	0.016	0.545	0.046	0.938	21
VE=80%	0.167	0.627	0.002	0.365	0.006	0.992	18



also less ambitious in its goals than the previous Phase 3 version by having looser Type 1 error control and no super-efficacy requirement. Even though the first proposal was formally called a "Phase 3 study" and did not have a dose selection part typically present in Phase 2/3, it was similar in spirit to a seamless Phase 2/3 design because the 1st part of the study (prior to the futility analysis) was playing a role of Phase 2. It was helpful to separate that part into a stand-alone Ph2b trial with its own objective in order to demonstrate proof of minimal efficacy in a select sub-population and also to size it appropriately to make a scientifically robust decision ($s = 26$ events rather than initial $s = 16$ which was too low). With this new simplified Ph2b design, a more ambitious claim of super-efficacy as well as multiple surgery type sub-populations would be addressed in a separate phase 3 study.

During team review and discussions, the proposed group sequential design was generally well-accepted. It was considered to have lower regulatory risk (at least in perception) because group-sequential methodology was considered "well-understood", while more complex adaptations such as unblinded sample size re-estimation were considered "less well understood" (FDA, 2010). The most important implication of this exercise was that such design consideration has prompted the team to formally evaluate operating characteristics of an early stopping decision rather than rely on some ad-hoc rules (as it was the case with initial Phase 3 proposal). Even though the newly proposed Phase 2 study was relatively modest in size (~2 K vs. ~10 K patients in the first proposal), it was still a huge investment. In such situations, teams are often asked to look for ways to accomplish more with that investment. So, during the internal review and discussions, the Phase 2/3 idea got re-introduced again, but this time with a different twist: the intention was to keep the general idea of the proposed Phase 2 design but make it more of a "pivotal" quality.

The discussion was prompted by the question whether early stopping for efficacy after 13 events should be part of the design. The study did not have efficacy stopping objective, but having a formal efficacy stopping boundary was considered beneficial

just in case the efficacy was so compelling that the independent data monitoring committee (iDMC) monitoring the study (for safety and futility) would recommend termination because of overwhelming efficacy. Although unlikely, such situation was not completely impossible and having a properly evaluated boundary with quantified operating characteristics was thought to be a better option than relying on iDMC's judgement alone. But once that possibility of early stop for efficacy was introduced, it raised the question: what should the sponsor do with the efficacy results if the study were terminated early for success? Would that data "stand on its own" in a filing package without the support of another large, expensive pivotal trial? The answer was: most definitely no, given very loose power/type 1 error requirements used for this study, not to mention too small of a sample size to provide sufficient safety information. That consideration lead to the team being asked to make the design "pivotal" quality while keeping its main elements. The main objective of the design revision was to be able to stop for efficacy early (in case the latter is overwhelming) and also make a super-efficacy claim at that point as well. The interim data resulting from such early stop would need to be of "pivotal quality" so that the study could be used in a filing package on its own, should the results look really compelling.

12.3.3 Semi-pivotal Phase 2b Followed by Pivotal Phase 3

With that new re-formulated objective, a new Phase 2b study was designed following the same procedure as outlined in the previous section. The following parameters/assumptions were used to design the new study:

- $\alpha = 0.025$ and 80% power for $\delta = 0$
- The primary focus of adding an interim look was futility stopping but, since data would have to be looked at in an unblinded fashion anyway, a formal boundary for efficacy look was included as well.
- An additional objective was added: if the minimal efficacy testing (Ho: VE <= 0 is rejected, then testing for super efficacy ($\delta = 0.30$) would be performed, with appropriate multiplicity adjustment via hierarchical testing.

The new study diagram is given in Fig. 12.4.

For this study design, a sample size of approximately $N = 2594$ patients would be required to achieve 42 total cases of *S aureus* infection. This number is approximate because in such setting, the actual sample size is a random variable whose expectation depends on underlying vaccine efficacy and background placebo rate. The $N = 2594$ was selected to balance primary and secondary objectives of this study and also to address the uncertainty about true underlying vaccine efficacy.

To achieve the primary objective of rejecting Ho: VE <= 0 (no super efficacy claim), $N = 2224$ subjects would be sufficient to guarantee 97.5% one-sided LCB to be above 0% with 80% power. This calculation was performed using the "base" assumption of VE = 60 and 3% placebo attack rate.

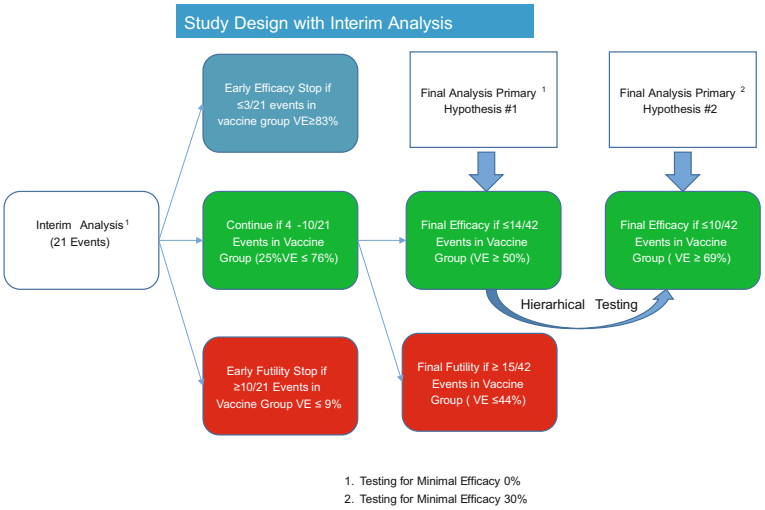


Fig. 12.4 Revised phase 2b study design with interim analysis

However, if slightly larger sample size of N = 2594 was used to achieve the same $s = 42$ events in case true vaccine efficacy was better than originally hypothesized, e.g. if VE = 80%, then the design would have 90% power to show 97.5% one-sided LCB > 30% (super-efficacy claim), assuming same placebo attack rate of 3%.

The highest of these two numbers was selected as enrollment target to balance both objectives.

Once the total sample size was fixed, an interim look for efficacy/futility was added after accrual of s = 21 events using group-sequential procedure described in the previous section. The resulting design is shown in Fig. 12.4. The operating characteristics of that design are shown in Table 12.6. These operating characteristics capture decisions related to the primary efficacy objective only (i.e. no super-efficacy claim). The probability of success/failure reported in that table describes the probability of demonstrating minimal efficacy.

12.4 Discussion

We have presented a story of clinical study design for vaccine efficacy study in a rare disease. Such studies are often complicated by very low background incidence rate of events in the population. The latter can be very uncertain and highly influential parameter which drives the sample size to be very large, even in cases of moderate to large treatment effect. Another factor driving large samples size is a super-efficacy requirement typical of vaccine studies: it is not enough to demonstrate that the vaccine is simply better than control (i.e.. VE > 0%) but it has to be better by certain amount. A

Table 12.6 Operating characteristics of final proposed Ph2b design

	Pr (Early eff. stop)	Pr (Early futility Stop)	PR (Final success)	PR (Final failure)	Overall prob of success	Ave # of events
VE = 0%	0.001	0.668	0.021	0.311	0.021	28.0
VE = 10%	0.002	0.576	0.044	0.379	0.046	29.9
VE = 20%	0.004	0.468	0.091	0.438	0.095	32.1
VE = 30%	0.008	0.349	0.179	0.464	0.187	34.5
VE = 50%	0.046	0.125	0.516	0.313	0.563	38.4
VE = 60%	0.109	0.050	0.687	0.154	0.797	38.7
VE = 70%	0.251	0.012	0.700	0.036	0.952	36.5
VE = 80%	0.527	0.001	0.470	0.002	0.997	30.9

Note

- Testing HO: VE $\leq$ 0% versus Ha: VE > 0%, assuming true VE = 60% and 1-sided $\alpha = 0.025$
- Decision Criteria (at interim analysis): Early Futility Stop if $\geq$10/21 Events in vaccine group; Early Efficacy Stop if $\leq$3/21 events in vaccine group
- Decision Criteria (at final analysis): Final efficacy if $\leq$14/42 Events in Vaccine; Final futility if $\geq$15/42 Events in Vaccine

non-zero null hypothesis, i.e.. $H_0 \leq 30\%$ or $H_0 \leq 40\%$ is often required for regulatory approvals of vaccine efficacy studies.

Two statistical approaches are commonly used in vaccine efficacy studies: modeling events via binomial distribution, i.e. 2-sample binomial problem or modeling events via Poisson distribution, then conditioning on total number events, thus reducing the problem to a one-sample binomial problem. This is often called "an event driven design".

The second approach is attractive (and more commonly used in practice) because it eliminates the dependency on the nuisance parameter (placebo event rate) by conditioning on the overall number of events in the power calculation. That is, the total number of events required to achieve desired type 1/2 error control does not depend on background event rate. However, the total number of patients will depend on the background rate. Therefore, this design should be interpreted with caution: the study size may look deceptively small (compared to 2-sample binomial design if used for the same problem), especially if one focuses on expected sample size without taking into account its variability. Even the expected sample size cannot be treated as a hard number, because it depends on vaccine efficacy-the very parameter we are trying to estimate. That is not to say though that this design should not be used: it can be useful and helpful as long as one realizes the variations in actual sample size enrollment and is prepared to handle them via increased complexity of study logistics (enrollment, flexible budget, timing etc.).

Regardless of which statistical approach is used, vaccine efficacy studies in rare disease present a unique challenge (due to uncertainty and high impact of assumption parameters) making them good candidates for exploring adaptive design options. In

such large trials, even small deviations from assumptions (VE and/or event rate lower than assumed) made at the design stage can lead to costly consequences. Clinical development program in such cases need to properly account for uncertainty about parameters of interest rather than just focusing on 2 values: Ve corresponding to the null and alternative hypotheses. In our study, the uncertainty about VE and λ_0 were only part of the problem. Other challenges included heterogeneity of population arising from multiple surgery types.

To tackle these challenges, the team went through 3 clinical development options:

1. (Extreme) Very complex "Phase 2/3-like" design attempting to address sub-population issue and uncertainty about parameters all at once via adaptation.
2. (Extreme) Very simple and small Phase 2b focusing on proof of minimal efficacy only and dropping sub-population and super-efficacy issues. Phase 3 would be a separate study.
3. (Middle-ground) Larger Phase 2b group-sequential trial keeping other elements of option 2 the same. The study was sized so that it could possibly be "pivotal" quality, should the data turn out to be really compelling. A separate Phase 3 study would include group-sequential elements (i.e. early futility stop) along with more complex adaptations such as sample size re-estimation.

The lesson learned from this story was that adaptive design cannot be used as a miracle answer in situations with competing multiple goals and high uncertainty about assumptions governing the study design: adaptive design is very unlikely to solve these problems at once, as we have seen with poor operating characteristics of option 1. By definition, adaptive designs are highly tailored tool to offer solutions for specific objectives. And to design a good adaptive study (i.e. having good operating characteristics and not just adaptive for the sake of being adaptive), one needs somewhat firm knowledge about underlying "truth" scenarios. Vaccine efficacy range of 0–100% as it was in option 1 was too ambitious of a goal to come up with adaptive design that would perform uniformly better than fixed design across the whole range of possible VE scenarios. In our consideration, it proved to be useful to break the problem into steps and assign priorities (i.e. to look at separate populations, to firm up VE and event rate before attempting to design a study with sample size re-estimation). This way the operating characteristics could be evaluated more thoroughly and the decisions were more transparent. In clinical program option #3 (which was selected) there still may be room for adaptation later (in Phase 3) but it will be done in a more thorough manner (unlike option 1) and after proper Phase 2b data readout. It is well known that unblinded sample size re-estimation works well only over narrow range of VE (close to alternative) and it does not eliminate "total" uncertainty about these parameters. In other words, one needs some data to design a good adaptive design, it can't be designed using guessed values only and perform well.

One final note on regulatory attitudes on this designs: the study was designed over 3 years ago when actual examples of adaptive designs in confirmatory settings were scarce. Significant progress has been made since then in that area, with some examples including vaccines case studies (Bauer et al. 2016; Lin et al. 2015). Even

though unblinded sample size re-estimation still remains classified as "less well understood" by draft FDA guidance (2010), it wouldn't be too much of an exaggeration to say that it is likely to change in the subsequent revisions of the guidance, based on recent experiences accumulated and regulatory attitudes "warming" up to such type of designs. Development programs like the one presented can be a good starting point for further incorporation of more sophisticated designs.

Acknowledgements Joseph Severs Deepthi Jayawardene, Yahong Peng, Scott Patterson, Neal Thomas, Qin Jiang, Christy Chuang-Stein.

References

Pfizer Inc. (2015, July 7). Pfizer begins phase 2b study of Its investigational multi-antigen staphylococcus aureus vaccine In adults undergoing elective spinal fusion surgery[Press release]. Retrieved from https://www.pfizer.com/news/press-release/press-release-detail/pfizer_begins_phase_2b_study_of_its_investigational_multi_antigen_staphylococcus_aureus_vaccine_in_adults_undergoing_elective_spinal_fusion_surgery.

ClinicalTrials.gov: https://clinicaltrials.gov/show/NCT02388165.

Nauta, J. (2010). *Statistics in clinical vaccine trials*. Berlin Heidelberg: Springer.

Jennison, C., & Turnbull, B. W. (2000). *Group-sequential methods with applications to clinical trials*. Boca Raton, London, New York, Washington DC: Chapman & Hall.

Anderson, K. (2016). gsDesign: Group Sequential Design. R package version 3.0-1. URL https://CRAN.R-project.org/package=gsDesign. Published 2016.

Bauer, P., Bretz, F., & Dragalin, V. (2016). Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine, 30,* 325–347.

Lin, Min, Lee, Shiowjen, Zhen, Boguang, Scott, John, Horne, Amelia, Solomon, Ghideon, et al. (2015). CBER's experience with adaptive design clinical trials. *Therapeutic Innovation & Regulatory Science, 50*(2), 195–203.

FDA. (2010). Draft guidance for industry—adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/…/Guidances/ucm201790.pdf. Published 2010.

Chapter 13
Patient-Reported Outcome Measures: Development and Psychometric Evaluation

Lori D. McLeod, Sheri E. Fehnel and Joseph C. Cappelleri

13.1 Introduction

This chapter has been created to provide an accessible introduction to the development and psychometric evaluation of patient-reported outcome (PRO) measures specifically designed to assess key endpoints in clinical trials, with the ultimate goal of supporting approval and/or labeling claims for pharmaceutical products. While many of our recommendations are broadly applicable to the development of PRO measures for use in clinical trials in any country and in other types of patient-based research (such as observational studies), this chapter will primarily focus on assembling and documenting the types of evidence needed to facilitate reviews of key study endpoints by the United States (US) Food and Drug Administration (FDA).

Figure 13.1 provides an overview of the steps described in this chapter and is adapted from the original "wheel-and-spokes" diagram provided in the FDA's Patient-Reported Outcome (PRO) Guidance, Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims (FDA 2009). The PRO Guidance and a subsequent document, The Roadmap to Patient-Focused Outcome Measurement in Clinical Trials (FDA 2013a), outline the information that the FDA recommends and reviews for both existing and newly created or modified measures used to support the assessment and computation of key study endpoints. This chapter provides an overview of the steps involved in the development and evaluation

L. D. McLeod (✉) · S. E. Fehnel
RTI Health Solutions, 3040 Cornwallis Road, Research Triangle Park, NC 27709, USA
e-mail: lmcleod@rti.org

S. E. Fehnel
e-mail: sfehnel@rti.org

J. C. Cappelleri
Pfizer Inc, Groton, CT 06340, USA
e-mail: joseph.c.cappelleri@pfizer.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_13

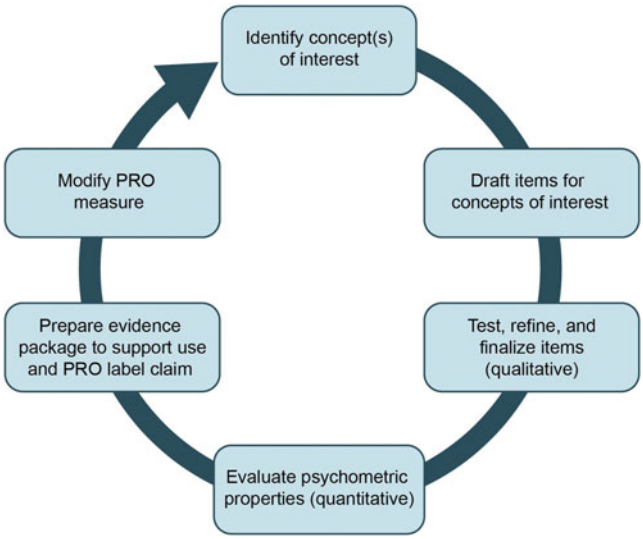


Fig. 13.1 Development, psychometric evaluation, and regulatory support of PRO measures (This figure provides a visualization of the process of developing and evaluating a patient-reported outcome (PRO) measure and submitting the measure for review. The figure is based on the original wheel-and-spokes outlined in the PRO Guidance (FDA 2009).)

of PRO measures, with a focus on meeting the review criteria described in each of these guidance documents. Rather than providing comprehensive instructions, we hope that this chapter will help readers learn to critically evaluate information about existing PRO measures and better understand the necessary steps when planning the development of a new PRO measure.

For more detailed descriptions of instrument development and psychometric evaluation processes, readers are encouraged to study full-length works by Nunnally and Bernstein (1994), Streiner et al. (2015), Fayers and Hays (2005), de Vet et al. (2011), Cappelleri et al. (2013), and Fayers and Machin (2016), as well as task force reports from the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) (Patrick et al. 2011a, b; Coons et al. 2009; Walton et al. 2015; Wild et al. 2005) and the International Society for Quality of Life Research (ISOQOL) (Reeve et al. 2013; Wyrwich et al. 2013) which are cited throughout the chapter.

While this chapter is focused on measures for which the assessment relies on patients' direct responses to a question or series of questions (i.e., patient-reported outcome measures), the steps and types of evidence described here are generally appropriate for the development and psychometric evaluation of other types of clinical outcome assessments (COAs) that may be used in clinical trials, including clinician-reported outcome (ClinRO), observer-reported outcome (ObsRO), and performance-based outcome (PerfO) measures (FDA 2013a; Cappelleri and Spielberg 2015). The FDA has provided a useful glossary for these and other COA terms

at https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm370262.htm.

Ultimately, the objective of this chapter is to describe the process of developing and documenting the psychometric properties of a PRO measure in order to ensure that the patient perspective is appropriately represented in the drug development process with the rigor necessary to ultimately support product approval or labeling claims.

13.1.1 Defining a Patient-Reported Outcome

The FDA defines a PRO as "any report of the status of a patient's health condition that comes directly from the patient without interpretation of the patient's response by a clinician or anyone else" (FDA 2009). This definition is similar to the European Medicines Agency's (EMA's) description, which was provided in a reflection paper on the Regulatory Guidance for the Use of Health-related Quality of Life Measures in the Evaluation of Medicinal Products (EMA 2005): "any outcome directly evaluated by the patient and based on patient's perception of a disease and its treatment(s)." A PRO measure that will be used to collect self-reported information from patients may be as simple as a single item or as complex as a multidimensional instrument.

13.2 PRO Measure Development

13.2.1 Concept Identification

The first step in the development of a PRO measure is to identify the concept(s) to be assessed, or **concept of interest** (Walton et al. 2015). Concepts selected for assessment in a PRO measure should be aspects of the illness or potential benefits of treatment that both are important to patients and have the potential for meaningful change within the context of a clinical trial. Importantly, there may be symptoms or impacts of a disease that are important to patients but are either not possible to measure within the context of a clinical trial or unlikely to change with treatment. While the collection of data pertaining to such concepts is important for understanding disease burden, key study endpoints in clinical trials should be based solely on measures with the potential to detect treatment benefits.

Examples of commonly assessed concepts of interest include signs and symptoms of the disease or condition, physical functioning, psychological well-being, activities of daily living, and health-related quality of life (HRQoL). While all of these concepts may be appropriate for measurement in clinical trials in order to demonstrate treatment benefits that are important to patients, product approvals and labeling claims based on PRO data in the US are more likely to relate to improvements in

symptoms which are proximal to the disease and the drug's mechanism of action (e.g., pain, itch) than those that are more distal and complex (e.g., HRQoL). For the purpose of product labeling, concepts that can be measured more objectively, such as symptom frequency and physical function, also tend to be favored over more subjective concepts, such as satisfaction and self-esteem (Gnanasakthy et al. 2012, 2017).

Patrick and colleagues (2011a) recommend that the overall **context of use** for the PRO measure and the resulting endpoint(s) be considered during the concept identification step. For example, even during the early stages of development, it can be extremely beneficial to determine how the measure fits within a preliminary endpoint model, a visualization which presents the hierarchy of endpoints to be tested within a clinical trial to support targeted labeling claims. Another potential exercise to consider during this step is the development of an overall disease model to help organize key aspects of the disease and treatment. For the purpose of this chapter, we focus on the identification of concepts that will provide the content for the items (questions) to be administered within the PRO measure.

Patients should generally be the primary sources of information for identifying concepts to be measured within a PRO measure; supportive sources may include the literature, clinical experts, and other stakeholders (e.g., caregivers, patient advocates), where appropriate. Concepts are generally elicited through qualitative research conducted with reasonably representative samples of patients drawn from the target patient population. Reviews of existing literature and instruments, as well as the solicitation of input from clinical experts, can help instrument developers identify important aspects of the disease and facilitate the collection of data from patients. Based on this information, an informed determination can be made as to whether it is preferable to adopt or modify an existing measure or to develop a new measure. If it is determined that a new measure should be developed, it is essential that the background information gathered at this early stage does not influence the qualitative research that forms the basis for the new measure.

13.2.1.1 Literature and Instrument Review

A review of medical literature in the relevant therapeutic area can provide a robust background for clinical aspects of the condition, symptoms and impacts that have already been identified, and existing measures that are available. In some cases, results of qualitative research or surveys may be available to help begin the process of identifying concepts that are important from the perspective of patients. Furthermore, the literature can provide information related to existing therapies, potential benefits of these therapies, common side effects, and the standard of care, including the role of parents or other caregivers (if relevant).

In addition to the published literature, sources such as meeting abstracts, clinical trials, product labels, practice guidelines, and any regulatory guidance documents in the therapeutic area of interest should be reviewed to identify concepts related to the target condition and its impact on patients. While never a substitute for the conduct

of methodologically rigorous qualitative research, social media can also be a source of relevant patient-reported information (Baldwin et al. 2011; Rothman et al. 2015).

As potentially important concepts begin to emerge, instruments addressing these concepts should also be reviewed. In addition to instruments identified from previously described sources, the FDA's COA Compendium (FDA 2016) is an excellent resource for identifying measures with the potential to support labeling claims. While inclusion of a PRO measure in this compendium does not guarantee a measure's acceptability for supporting labeling claims (and, similarly, exclusion does not preclude acceptability), the FDA has provided this very useful resource to help sponsors identify PRO measures and other COAs that the FDA is willing to consider and discuss in a wide variety of therapeutic areas; often these measures have been used previously to support labeling claims. There are also instrument databases that can be reviewed to identify potentially relevant measures as well as information regarding the development of these measures and evidence supporting their use, such as the Patient-Reported Outcome and Quality of Life Instruments Database (PROQOLID; https://eprovide.mapi-trust.org/); Patient-Reported Health Instruments (PHI; http://phi.uhce.ox.ac.uk/home.php), a database maintained by the University of Oxford; and the Measurement Instrument Database for the Social Sciences (MIDSS; http://www.midss.org/). Instrument databases and compendia are also available in various therapeutic areas.

13.2.1.2 Clinical Expert Input

While patient input reigns supreme in the development of a PRO measure, input from experts in the therapeutic area or condition under study can be extremely useful throughout the development process. For example, clinical experts can facilitate concept identification by providing additional background about the condition, offering insight regarding key symptoms and impacts based on interactions with patients, and identifying unmet needs and desirable treatment attributes based on their experience.

The optimal number of clinical experts and the extent of their involvement varies based on the degree of clinical expertise within the instrument development team, the complexity/nature of the PRO measure, and logistical concerns such as budget and timeline. Commonly, the involvement of clinical experts begins with individual interviews or an expert panel meeting to gather pertinent background information and identify patient-reported concepts that are important from the perspective of clinicians, as well as any potentially relevant PRO measures with which the experts may be familiar. Additional input solicited during the course of the study may relate to plans for the qualitative research, including input or feedback on the screening criteria and interview/focus group guide; review of qualitative results to identify clinically relevant concepts with the potential for change within the context of a clinical trial; and clinical guidance pertinent to the psychometric evaluation process, such as the assessment of clinical characteristics and input into the study design.

13.2.1.3 Direct Patient Input

While reviews of existing information and guidance from clinical experts can contribute information to the concept identification process and facilitate the conduct of patient-based research, the most important step in identifying concepts for measurement is the elicitation of input directly from patients.

Direct patient input is generally gathered during semi-structured qualitative interviews; participants are asked a series of open-ended questions aimed at furthering the researchers' understanding of the disease, identifying concepts important to patients, and observing patients' word choice when describing these concepts to inform item development. For example, when developing a symptom-based measure, interview participants will be informed that the goal of the interview is to identify a comprehensive set of symptoms and to fully understand how the patient experiences and perceives each of these symptoms. Participants are then asked to describe each of their symptoms and encouraged to be as explicit and inclusive as possible.

Additional questions are generally posed to gather specific information that may not be mentioned spontaneously, such as the frequency, duration, and severity of each symptom. In order to hear the words patients use to describe improvement and worsening of their symptoms and to inform the selection of an appropriate recall period for the PRO measure, the variability in the experience of each symptom (e.g., during the day, from day-to-day) is also queried. Symptoms that are not reported spontaneously but that have been identified as important or common in the literature or based on clinician input will also be explored to determine the potential relevance and importance of these concepts to patients.

Generally speaking, symptoms that patients mention spontaneously tend to be the most salient compared to those symptoms that are only endorsed in response to follow-up questions; as such, spontaneously reported symptoms may be the most important targets for measurement. Additional techniques, such as importance rating and symptom ranking, may also be used to further elucidate the relative importance of concepts elicited from patients.

While group discussions, often referred to as focus groups, offer a reasonable and, in some regards, more efficient alternative to patient interviews (i.e., permit the simultaneous solicitation of information from multiple patients), individual interviews are generally preferred for the purpose of concept elicitation, as this setting allows each individual to provide detailed information about his or her own experiences without the influence of others. Individual interviews are particularly valuable when working with special populations (e.g., children, sensory impaired, very ill) or discussing sensitive topics, and they allow greater flexibility to accommodate participants' schedules.

Regardless of the qualitative data collection method chosen, participants should be selected based on criteria that are consistent with the future clinical trial screening criteria and provide a broad representation of patients in terms of relevant demographic variables (e.g., age, gender, race) and clinical characteristics (e.g., subtypes of disease, varying levels of symptom severity). While patients receiving efficacious treatment are typically excluded from clinical trials, inclusion of such patients in

qualitative research can be beneficial, particularly when discussing the meaningfulness of treatment-related changes.

For prevalent conditions in which screening criteria can be self-reported, the identification of patients is often possible through qualitative research firms (commonly called focus group facilities), many of which employ medical recruiters and have extensive databases that include health-related information. However, the involvement of clinical sites may be necessary if the screening criteria are complex or require information that potential participants may not have readily available (e.g., test results). For rare conditions, recruitment through clinical sites and/or patient advocacy organizations is often necessary.

The number of patients needed for the concept identification stage varies depending on the amount known about the condition, the variability of disease characteristics and symptoms, the number of subgroups of interest, and the potential complexity of the PRO measure. For example, the elicitation of concepts for a measure being developed to address the severity of a single symptom experienced by all (or nearly all) patients with a given disease will require a smaller sample size than a measure being developed to assess HRQoL within a diverse patient population. As a general rule, the concept elicitation process should continue until **concept saturation**, the point at which no new information is being elicited, is reached. The achievement of concept saturation should also be documented by developing a detailed table that shows the concepts elicited in each interview (or focus group) and the number of new concepts elicited in later interviews compared with earlier ones.

To ensure that the qualitative research process is appropriately documented and systematic, a qualitative research protocol (including recruitment and enrollment details, an informed consent form, and an interview guide to help structure the discussions) should be developed prior to initiating patient interactions. This document should be approved by the appropriate institutional review board (IBR) or ethics review committee before any patients are recruited. Throughout the development process, it is important to create transcripts to document the concepts of interest as described by the patients in their own words. Having two members of the project team with qualitative research experience is also recommended; this allows one interviewer to take the primary role of facilitating the interview or focus group while the other takes field notes that will be supplemented by the transcripts for later data analysis and documentation. Careful documentation is particularly important, since the FDA commonly requests the opportunity to review both the qualitative research protocol and the transcripts when reviewing requests for product approvals and labeling claims based on PRO measures.

13.2.1.4 Finalization of Concepts

Using both transcripts and field notes, the PRO measure development team reviews the results of the interviews. Terms, phrases, and statements made by participants are generally coded to facilitate analysis of the qualitative data. While such a coding scheme is generally constructed prior to analysis, additional codes are commonly

added during the analysis process as new concepts are identified. The coded results are then compared across interviews or focus groups (or both) to generate themes or patterns in the way participants describe the concepts, as well as to identify the concepts reported most commonly and deemed most important by participants.

Software such as Atlas (www.atlasti.com) can be used to facilitate the analysis process by organizing the data according to assigned codes but does not replace the need for experienced qualitative researchers to synthesize the information. In addition to detailed descriptions of the results, tables are typically developed to summarize the concept elicitation results, support the development of items, and document concept saturation. While all concepts deemed important by patients and relevant to the goals of the PRO measure are potential candidates for measurement, the input of clinical experts can be very valuable at this stage to ensure that each of the concepts selected for item development is clinically relevant within the context of use. For example, when developing a symptom-based measure, the symptoms selected for measurement should be plausibly related to the condition under study from a clinical or physiological standpoint.

13.2.2 Item Development

Once the concept or set of concepts to be assessed by the PRO measure has been identified, items should be drafted to address each concept of interest using patient-friendly language and methodological principles grounded in survey research (Sudman and Bradburn 1982); the focus of the question-writing process should be on facilitating patient understanding, yielding proper response, and minimizing measurement error. For example, questions should be succinct and based on the language used by patients, maintain an accessible reading level, and naturally relate to the response options. Item developers should be careful to avoid "double-barreled" questions or items that address more than one concept that might have different or conflicting answers. For example, an item may ask if a person has trouble walking or jogging. When answering this item, an individual who can walk but not jog must determine which component to answer and which to ignore.

Response options should be both comprehensive and mutually exclusive; it is also important that patients be able to clearly differentiate among each option. Careful attention must also be given to the recall period to ensure that patients are able to accurately remember relevant information and formulate an accurate response without complex computations. While very short recall periods (e.g., past 24 h) and frequent administration have been recommended by the FDA for assessing symptoms with the potential to vary from day to day, there may be situations in which a longer recall period and less frequent administration are more appropriate (Norquist et al. 2012). Longer recall periods may be appropriate for the assessment of symptoms that are slow to change, for example, or for determining the impact of chronic conditions on patients' functioning and HRQOL.

When developing items for a PRO measure to be administered in clinical trials, it is also vitally important to consider the context. Specifically, these measures need to be designed to evaluate treatment benefit. As such, patients' experiences (and, consequently, their responses to the PRO items) must have the potential to change over the course and within the context of a clinical trial. The items and the PRO measure itself should be brief in order to minimize patient burden and allow for flexibility in the mode of administration. For example, a PRO measure being developed for electronic administration on a daily basis should not include more items than patients will be willing to answer consistently and the individual items must be brief enough to display clearly on a small screen (e.g., hand-held device or smart phone). Finally, items should be applicable to and generalizable across the target patient population.

Multiple items are typically drafted for each concept of interest so that variable question wording and response scales (e.g., numerical rating scales, verbal response scales) can be tested in cognitive debriefing interviews to identify the version of the item that is most easily understood and answered by patients. It is also helpful to consider potential scoring rules during item development. For example, including a category for "no pain" can be justified as the lowest (most favorable) score for a pain severity item; however, if a "not applicable" response option is included, this response may be chosen by patients who did not experience pain or by patients who restricted their activity to avoid pain, creating measurement error in the item score. Instructions for completion of the PRO measure should also be drafted at this step so that patient input can be gathered to refine instructions and maximize comprehension of the item(s) as an organized unit.

13.2.3 Cognitive Debriefing

Draft versions of the instructions and items are refined based on additional patient input gathered during iterative sets of interviews, commonly called cognitive interviews. In addition to cognitive debriefing of the PRO measure, these interviews typically involve the collection of supplemental data to further support content validity and inform future use of the final measure. Additional information on the conduct and analysis of data collected during cognitive interviews can be found in Willis (2005, 2015), respectively.

In general, cognitive debriefing focuses on the cognitive processing involved when participants read, interpret, and determine their responses to the draft questions. The insights gained through evaluating this processing step are then used to refine instructions, question wording, response categories, formatting, and other aspects of the PRO measure to remove aspects that are unclear or influence participants to understand the items in a way that was not intended. The process is iterative and should continue until evidence is established that no further revisions are warranted.

Typically, the cognitive debriefing sample is independent of the concept elicitation sample, although some overlap may be reasonable, particularly when working in rare diseases. Similarly, it is best practice to conduct cognitive interviews in person to

facilitate item review and behavioral observation; however, web-based or telephone interviews may be needed in some circumstances. While the patient sample should also be reasonably representative of the expected clinical trial participants, those with lower education may be oversampled to ensure that items are easily understood with lower levels of literacy.

As with the concept elicitation interviews or focus groups, a semi-structured interview guide should be developed to ensure consistency across the cognitive interviews. Commonly, the guide (and the interview) begins with a brief concept elicitation section to identify concepts important to these additional patients prior to reviewing the draft items; congruence among the concepts identified in the prior interviews, addressed in the draft PRO measure, and reported by cognitive interview participants further supports the content validity of the final measure.

During cognitive debriefing, participants are generally asked to "think aloud," describing their thought processes as they read, interpret, recall necessary information, and respond to the draft items. It is often helpful to provide an example and explain to participants that it is important for interviewers to hear this process across patients to identify any differences in interpretation or difficulties associated with recall or response which may suggest revision of the items. If not offered during the think-aloud process, participants are typically asked to paraphrase the question in their own words and to describe how they arrived at their response. A series of additional questions follows (as needed) to gather further information to fully elucidate the question-answering process. Patrick and colleagues (2011b) describe the goals of these interviews as two-fold: (1) to ensure that the most important concepts are included in the final PRO measure; and (2) to ensure that respondents understand how to answer each item based on clear instructions, the appropriate recall period, item meaning, response scale, and any other features, such as paper versus electronic mode of administration.

As mentioned previously, it is important to take the time to assess the clarity and appropriateness of instructions, question wording, response categories, and aspects of formatting with each individual to guide the finalization of the PRO measure. In addition, if subscales are desired for specific concepts of interest, gaining an understanding of how patients perceive relationships among the items will help inform the development of initial scoring algorithms. It is also useful to pose specific questions to ascertain what amount of change on the individual items and sets of items is meaningful to patients in order to facilitate the development of responder definitions for application in future clinical trials.

As with the concept elicitation step, the number of patients needed for cognitive debriefing will vary depending on the complexity of the concepts to be measured and variability across the patient population. In general, however, three or four iterative sets of interviews each comprised of six to 10 patients (for a total of 18–40 additional patients) are typically sufficient to help refine and finalize the items. As a general rule, the cognitive interviews are complete when no additional item changes are identified by subsequent interviews.

As with the concept elicitation interviews, cognitive interviews should be guided by a brief protocol. IRB or ethics approval must be obtained prior to patient engage-

ment, and transcripts should also be created to support revisions and document this step of development.

Immediately following each cognitive interview, the researchers conducting the interviews begin the refinement process by reviewing the field notes and discussing the patient feedback for each item. Generally speaking, the modifications made between sets of interviews will be based on the field notes to meet development timelines. However, these initial findings will be followed by a more formal analysis facilitated by interview transcripts from which patient quotes are identified that best identify and support the refinements needed.

An **item tracking matrix**, a log that provides the chronology of events for item generation, modification, and finalization, is typically created to document the refinement process and rationale for modifications. This item tracking matrix is commonly requested by FDA reviewers and describes the item wording for each version of each item tested, reasons for revisions to retained items, and reasons for the omission of items during the process. The item tracking matrix also provides evidence that further refinements are not needed by documenting subsequent interviews in which there was no feedback identifying weaknesses or suggesting revisions to clarify the question wording or response choices.

13.2.4 Potential Exceptions and Additional Considerations

It should be noted that the steps outlined in the preceding sections are those recommended for the "typical" instrument development process (if, indeed, there is such a thing). There are many instances in which alterations or additions to these steps are either necessary or advisable. For example, if a PRO measure is meant to address a single concept (such as the frequency of a particular symptom) or if a great deal is already known about the concepts that need to be measured based on prior research, it may be reasonable to combine the concept elicitation and cognitive debriefing steps. Specifically, items may be drafted and tested as described in Sects. 13.2.2 and 13.2.3. In that case, however, the interviews would begin with a comprehensive concept elicitation phase to support the content validity of the final measure in addition to refining the item set. Such an approach may also be appropriate if one is modifying an existing PRO measure rather than developing a new measure.

Just as the nature and mode of administration need to take into consideration the characteristics of the target patient population, the development process often needs to be tailored when working with special populations, such as children and individuals with cognitive or sensory impairments. Recommendations for the conduct of qualitative research in such patient populations have been provided by DeMuro and her colleagues (2012).

If a PRO measure is meant to be administered in international trials, the development process should ideally be completed in multiple countries. Much like qualitative research, samples should be representative of the patient population in terms of demographic and clinical characteristics; development of a PRO measure involving

patients from different countries allows for diversity in geographical and cultural characteristics. It is very common, however, for the initial development of a PRO measure to be limited to a single country. In such cases, translation and cultural adaptation are then performed later, in preparation for administration in international studies.

When instrument development takes place in a single country but the measure is expected to be used in international trials, it is often useful to conduct a translatability assessment. This assessment includes an evaluation of the draft items by translation experts to identify any potential issues with the item wording (language or concepts) and response options/scales. The objectives of this assessment are to increase the likelihood of cultural equivalence across languages and reduce the potential for translation issues caused by items or phrases that are difficult to translate. Translatability assessments should be conducted during the item refinement process so that modifications suggested by translators can be tested with additional cognitive debriefing participants. The recommended methodology for these activities has been detailed by Wild and her colleagues (2005).

Considerations related to the development of a PRO measure for electronic administration (ePRO) are analogous to those pertaining to development in multiple languages. Ideally, the development process involves cognitive debriefing of ePRO versions of the items with patients using the device slated for use in the clinical trials. This process negates the need to demonstrate measurement equivalence between different modes of administration (e.g., ePRO with paper), because the ePRO version is the original and only version of the measure. However, development of a paper-based PRO measure which is later migrated to an electronic platform or migration from one type of electronic platform to another (e.g., web-based to hand-held) is not unusual. Much like the translation and cultural adaptation processes, ePRO migration must follow a rigorous process to ensure that the content validity and psychometric properties of the final version are maintained. The evidence required to support measurement equivalence between paper-based and ePRO versions of PRO measures has been described by Coons and his colleagues (2009).

13.2.5 Documentation of the Development Process

Whether separately or in a comprehensive development report, it is important to document all steps involved in the literature review, expert involvement, qualitative research for concept elicitation, final concept selection, item development, and item refinement. In addition, a conceptual framework for the new measure should be developed. The conceptual framework depicts the concept that each item (potentially in conjunction with other items) is meant to assess in the version of the PRO measure being taken forward for psychometric evaluation. This framework could be as simple as a single item addressing the severity of a single symptom or as complex as groupings of items underlying various domains of HRQoL. The conceptual framework provides an initial glimpse into possible scoring for the PRO measure based on

the development phase and may be refined based on the results of the quantitative phase in which the relationships among item scores are evaluated empirically. These documents will serve as the basis for regulatory packages to facilitate review and agreement that the PRO measure is fit for purpose—that it is an appropriate measure in the proposed context of use for supporting medical product labeling.

13.3 Psychometric Evaluation

The PRO guidance (FDA 2009) describes the important psychometric properties that must be demonstrated for PRO measures used in pivotal clinical trials in order to support product approvals and labeling claims in the US. Specifically, the guidance describes the evaluation of the reliability, validity, and ability to detect change (often referred to as responsiveness) as the psychometric groundwork for the PRO measure. In general, the quantitative assessment of these measurement properties (the "validation") should be made within the same context of use as planned for the pivotal clinical trials to support labeling claims.

At the most basic level, the purpose of the psychometric evaluation is to gather evidence that the PRO measure is reliable and valid within the intended context of use. Generally, the same properties should be evaluated regardless of the intent of the PRO measure or whether the PRO measure is newly developed or an existing measure being used in a different condition or for a different purpose. For a PRO measure that is intended to support product approval or labeling, the amount of evidence required is substantial in order to ensure that the concepts important to patients are reliably and accurately captured and that the measure is capable of detecting treatment benefit. For the evaluation of an existing PRO measure, the evidence required will include that the measure's psychometric properties indicate that the scores behave adequately in the new context. In the next sections, each of the psychometric properties is defined and a method for evaluating the property is described. Rather than being exhaustive, the methods provided here are intended to be examples of how each property is commonly assessed. For more detailed information on psychometric evaluations, readers are encouraged to consult the in-depth literature recommended at the beginning of this chapter as well as consortium recommendations (Frost et al. 2007; Reeve et al. 2013) and publically available summaries based on FDA review (FDA 2013b). Although we describe approaches that are recognized as best practices for conducting psychometric evaluations, practical constraints should be acknowledged and considered, especially for conditions such as rare diseases, where the FDA often shows flexibility in terms of the sample sizes, number of studies, and amount of supportive evidence required.

13.3.1 Psychometric Evaluation Data

Preliminary psychometric evaluations may be planned using cross-sectional data outside of the clinical trial program, especially if additional item reduction is planned. However, evaluation through cross-sectional data is not sufficient to assess all of the key properties. If a PRO measure has been developed rigorously and substantial item reduction is not planned, the most efficient way to gather data for a comprehensive psychometric evaluation is to administer the PRO measure as an exploratory measure within a planned phase 2 study. When planning the phase 2 study, it is important to include additional measures that can support the psychometric evaluation (i.e., measures of similar constructs and measures that provide a global assessment of disease status and/or of change in disease status from baseline). The PRO measure should be administered at key time points aligned to the intended time points for future pivotal studies. These time points should provide opportunities to evaluate change in the proposed scores where change is anticipated and to evaluate stability of scores where change is not expected. The number of additional measures and the frequency of their administration should be taken into consideration, however, to avoid increasing patient burden unnecessarily. Finally, blinded phase 2 study data should be used for psychometric evaluation of the PRO measure.

If phase 2 data are not available for the longitudinal evaluation, an independent study may be conducted to collect appropriate data for the psychometric evaluation. In such cases, it is important to ensure that the sample is representative of the intended pivotal studies and that the study design facilitates the collection of data at time points where change is expected (e.g., prior to and after treatment) and where stability is expected for at least a subgroup of patients (e.g., prior to treatment or after treatment benefit has been maximized). In addition, it is important that the selected study has adequate sample size for the methods planned (Chen et al. 2014).

13.3.2 Psychometric Properties

The psychometric evaluation should be guided by a formal psychometric analysis plan. If the psychometric evaluation is performed using phase 2 data, the psychometric analysis plan should be a separate document from the statistical analysis plan designed to guide the evaluation of efficacy based on the phase 2 data. The psychometric analysis plan should describe the psychometric and statistical methods to be conducted (including prespecified analytic sample and subgroups of interest), and provide detailed hypotheses to support reliability and construct validity as well as the ability to detect known differences between distinct groups, and meaningful change for each score derived from the PRO measure. Blinded data should be used. Table 13.1 provides an overview of typical analysis methods for each of the key psychometric properties.

Table 13.1 Key psychometric properties assessed

Psychometric property	Measures used: examples	Evaluation method: examples
Distributional characteristics		
	The target PRO measure item-level and scale-level scores at key time points. Scores on supporting measures (e.g., measures intended for construct validity and known-groups comparisons) at key time points	Mean, median, SD, minimum, maximum, percentage missing, frequency distribution
Structure		
	Item-level scores at baseline and/or follow-up time points	Factor analysis
Reliability		
Internal consistency	Item-level scores at key time points	Cronbach's coefficient alpha
Test-retest	Item- and/or scale-level scores at two time points for a subgroup of patients whose disease status should be stable	Weighted kappa; Intraclass correlation coefficient
Validity		
Construct	Item- and scale-level scores for the target measure and supportive measures at multiple time points	Correlation
Known groups	Classification of patients into groups (for which PRO scores are expected to differ) based on an external variable at one or multiple time points	Analysis of variance or nonparametric test comparing item- and/or scale-level scores for these groups
Ability to detect change		
	Item- and scale-level change scores for the target measure and supportive measures	Correlation, effect size, standardized response mean, responsiveness statistics
Interpretation of change scores		
	Candidate anchor measure scores at multiple time points	Descriptive, correlation, regression, and cumulative distribution function plots
	PRO measure scores at baseline	One-half SD of the PRO measure scores at baseline

Note SD = standard deviation; PRO = patient-reported outcome

13.3.2.1 Distributional Characteristics

Key demographic and sample characteristics at baseline should be tabulated to describe the sample used in the psychometric evaluation. Standard descriptive statistics (including means, medians, standard deviations, minimums, maximums, and percentages missing) should be reported for the PRO measure scores as well as the supporting measures included for comparison at baseline and key time points. Item-level frequencies should also be tabulated for these time points.

Review of the distributional characteristics provides a general evaluation of compliance and appropriateness of the PRO measure. Evidence of good compliance (i.e., minimal missing data across items and time points) provides support that patients are capable of completing the measure and that the patient burden of completion is acceptable. Scale- and item-level appropriateness are judged by the distribution of response categories across time points. Supportive evidence is defined by distributions where all categories are selected by at least a small proportion of subjects without evidence of floor and ceiling effects.

For the present chapter, a floor effect is defined by a high proportion of subjects selecting the minimum category or score for the measured construct (e.g., more than 80% select the minimum category). Alternatively, a ceiling effect is defined by a high proportion of subjects selecting the maximum category or score (e.g., more than 80% select the maximum category). Both floor and ceiling effects can indicate that the PRO measure may not have the measurement range necessary to differentiate and show change. For example, if a majority of subjects select the extreme category for low symptom severity at baseline, it is plausible that the symptom measured is not relevant for the majority of the intended population. Careful review of the concept addressed by the item and the characteristics of subjects who endorse the extreme category should be examined to more accurately determine whether this item is only applicable to the most severe patients. This exploration of potential reasons for the observed floor effect will inform decisions regarding whether to remove, revise, or retain this item. Ceiling effects are of less concern at baseline when efficacious treatment should facilitate lower (here defined as more favorable) item scores.

Evidence that the distributions for the supporting measures behave as expected provide an initial quantitative context for the target PRO measure. For example, if there is little variability in both the target PRO measure and in a related supportive measure, then the concern may shift to the relevance of the sample or overall context of use. On the other hand, if the supporting measures have the expected distributions but the target PRO measure does not—especially during post-baseline assessments with a beneficial intervention—the initial descriptive statistics for the target PRO measure may foreshadow problems with other properties.

13.3.2.2 Structure

For multiple items with the potential for subscales, the conceptual framework provides a description of the item-level relationship based on the qualitative phase of

the development process and how the items form potential domains. To inform scoring, this framework is further refined by evaluating the quantitative relationships among the items through inter-item correlation coefficients and dimensionality analysis (e.g., factor analysis). Methods such as exploratory factor analysis (EFA) are recommended when these frameworks are preliminary and alternative item groupings have been considered (e.g., consideration of one general grouping including all items versus grouping of item subsets into multiple domains). EFA seeks to "explore" the underlying factor structure of a set of variables (i.e., item scores) and therefore does not impose a structure. In this type of analysis, the inter-item relationships are generally used to evaluate the number of possible factors representing constructs or domains for which subscale scores may be warranted. Items are allowed to freely load on the factors (that is, correlate with other items so that they group to form factors) based on relationships that meet prespecified fit criteria. Multiple structures (e.g., 3 factors, 4 factors) are then compared in terms of factor loadings, variance explained by a factor, and overall model fit. The "best" structure is one where each item loads (correlates) highly (e.g., factor loading of 0.4 or above) on one and only one factor. Different EFA estimation techniques are recommended depending on the type of item response and the hypothesized relations among the items (Gorsuch 1983). Methods may vary based on, for example, rules for extracting (or forming) factors or determining the optimal number of factors, or based on the correlation allowed among factors (rotation).

Alternatively, confirmatory factor analysis (CFA) facilitates the evaluation of a proposed structure. In this type of analysis, the number of factors and the items that should load on each factor are prespecified. Items are grouped based on previous qualitative or quantitative results to form hypothesized domains. In addition to the groupings, the magnitude of the loadings and correlation among factors can be prespecified. For structures that are nested (i.e., one structure is a modification of another through the elimination of specific loadings or paths between factors), comparisons are facilitated through a likelihood ratio test of model fit. The general fit of the model can also be assessed using a variety of fit indices, such as the Comparative Fit Index (Bentler 1989); the Non-Normed Fit Index (Tucker and Lewis 1973); and a root-mean square error of approximation (Browne and Cudeck 1993). Figure 13.2 provides an example of the parameters estimated for an EFA for a 7-item measure; Fig. 13.3 presents a CFA where two factors are proposed a priori.

Results from the inter-item correlations and factor analysis can be used to propose or refine the preliminary domain construct for the PRO measure (i.e., which item scores are combined to form domain-level scores and to identify potential redundancies).

In addition to factor analysis, methods such as Rasch modeling and item response theory (IRT) modeling are useful tools to evaluate the performance of individual items and the structure of a multi-item PRO measure (Andrich 1988; Edelen and Reeve 2007; Cappelleri et al. 2014). These methods can be used to assess the relationships among items, how the items relate to an underlying construct, and redundancy in item content, making the methods extremely informative for item reduction and subscale (domain-level) refinement. Related methods, such as differential item functioning,

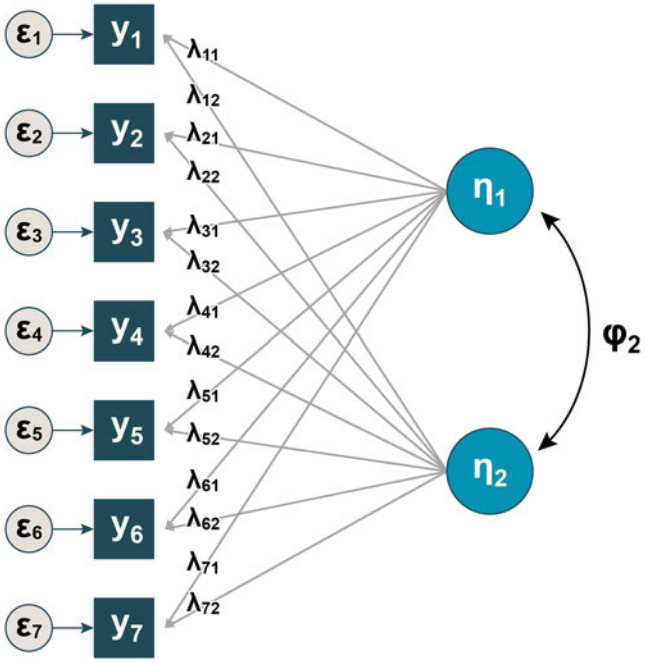


Fig. 13.2 Example exploratory factor analysis path diagram (Example exploratory factor analysis diagram for a 7-item measure. In this analysis, a 2-factor solution is fit with each item allowed to load on each factor. The correlation between the factors and a uniqueness for each item are also modeled. Note: y represents the observed item score. ε is the item's unique or error component that is not correlated with the factors. λ is the item's factor loading. η is the factor score. Φ is the factor correlation.)

are beyond the scope of this chapter but can be used to statistically test item scores for the presence of bias that is not related to the intended construct (e.g., gender or ethnicity bias in scores) (Thissen et al. 1993). Item-level results can be used to inform item selection or item revision while retaining the intended content. Furthermore, at this step it is important to consider scoring for multi-item scales. A simple score should be selected over a complicated score when the necessary precision is not compromised. In addition, when possible, reporting scale-level scores on the same scale as item-level scores will provide context for the scale-level interpretations.

13.3.2.3 Reliability

Reliability is defined as the extent to which an instrument is free of measurement error and can consistently measure a subject's true score, which is the average score that would be expected if the subject completed parallel forms of the instrument many times (Hays and Revicki 2005). It is important to evaluate the reliability of

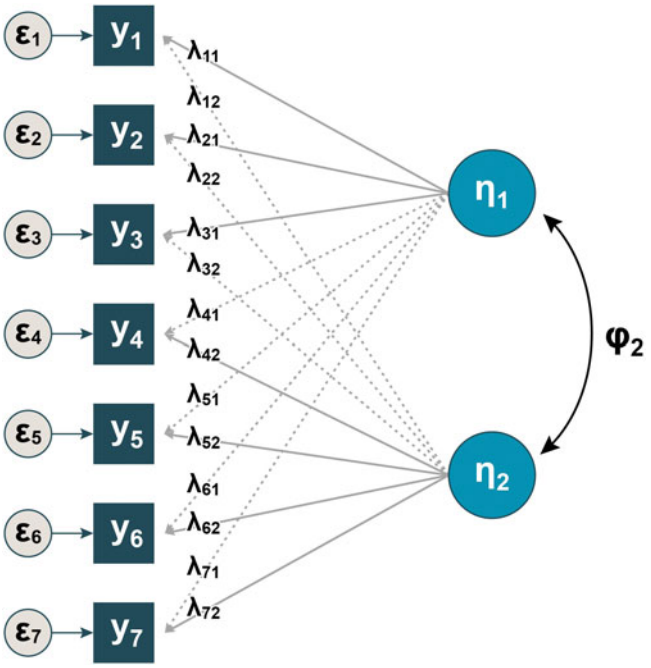


Fig. 13.3 Example confirmatory factor analysis diagram (Example confirmatory factor analysis diagram for a 7-item measure. In this analysis, a 2-factor solution is fit with specific items allowed to load on one and only one factor (and the specific factor is predefined). These loadings are identified by the solid lines and those not fit are identified by the dashed lines. The correlation between the factors and a uniqueness for each item are also modeled. Note: *y* represents the observed item score. ε is the item's unique or error component that is not correlated with the factors. λ is the item's factor loading. η is the factor score. Φ is the factor correlation.)

a PRO measure in order to provide evidence for the reproducibility of its reported scores. Two types of reliability assessments are recommended in the FDA's PRO guidance (2009): internal consistency and test-retest reliability.

Internal Consistency Reliability Internal consistency reliability explores the degree to which the items constituting a scale or subscale are associated and their ability to measure a single underlying construct. If items are highly related to one another, then less unassociated measurement error is present. Internal consistency reliability is typically estimated for each group of items proposed to form a single score using Cronbach's (1951) coefficient alpha, with alpha values between 0.70 and 0.95 providing evidence in support of adequate internal consistency reliability for outcomes measures designed for group-level comparisons. (Alpha higher than 0.90 suggests item-level redundancy and may not be desirable when balancing score precision and patient burden.) For measures intended to summarize distinct aspects of a condition that may not be related (e.g., differing symptoms that may not typically present together but are indicative of a condition), internal consistency is not

an appropriate type of reliability to evaluate, as these items are not indicators of the same underlying construct.

Test-Retest Reliability Test-retest reliability provides an evaluation of reliability by comparing scores for subjects who are classified as stable in the measured constructs across two time periods where no change is anticipated. For example, the subgroup may be defined as the group of patients who have been assigned to placebo and use time points where the placebo effect has resolved. Alternatively, the subgroup may be defined at the end of the treatment period using the pooled sample where additional treatment-based improvements are not anticipated. More commonly, stable subgroups are predefined as exhibiting no change based on a criterion measure that is judged to be able to adequately assess patients' status. Intraclass correlation coefficients (ICCs) are then computed using the scores for the "stable" subgroup at the two time points (Deyo et al. 1991; McGraw and Wong 1996; Shuck 2004). The "test" data are scores at the first of these two time points and the "retest" data are scores at the second time point. The ICC is the ratio of the between-subject variability and total variability; if this ratio is high, there is little measurement error to be accounted for by the measurements on the two different occasions, and the reproducibility (i.e., test-retest reliability) of the measure is high. ICC values of at least 0.70 for multi-item scales are recommended to support adequate test-retest (e.g., Nunnally and Bernstein 1994). For single items, methods such as ICCs and weighted kappa coefficients provide information related to test-retest reliability.

13.3.2.4 Construct Validity

Validity is traditionally defined as the extent to which an assessment measures what it purports to measure, but the term is also more broadly intended to convey the appropriateness of inferences based on item scores, subscale or domain scores, or total scores (Messick 1989). Within the context of PRO measures, regulatory bodies have focused primarily on content validity (as described in the Sect. 13.2). Without adequate evidence to support content validity, most regulatory reviewers will dismiss a measure from further review of the other aspects of validity. These additional aspects of validity use quantitative data and methods to build upon the qualitative evidence to further support the content validity. Construct validity describes the relationships among multiple indicators of a construct and the degree to which the scores on these indicators follow predictable patterns.

Convergent/Divergent Validity Technically speaking, convergent/divergent validity describes the relationships among multiple indicators of constructs and the degree to which the scores from these indicators follow predictable patterns. The goal of these analyses is to demonstrate stronger relationships among measures addressing similar constructs (defined as "convergent" validity) compared to measures addressing more disparate constructs (defined as "divergent" validity). Typically, correlational analyses are conducted to examine these relationships using data from baseline

and other key time points. Specifically, the psychometric analysis plan should outline a priori hypotheses that specify the anticipated direction and magnitude of correlations between scores on the target PRO measure and scores on other supporting measures included in the trial (i.e., other measures of both similar and different constructs, relevant clinical measures). Correlation analyses are then conducted and the patterns are compared against the a priori hypotheses.

Known-Groups Validity Known-groups validity evaluations focus on whether scores on the PRO measure differentiate (or "discriminate") among subgroups that should have different scores if the PRO measure is performing as intended. Again, these assessments begin with a priori hypotheses that are then evaluated. Analyses of variance or nonparametric tests can be used to examine mean differences in the PRO measure scores for the subgroups. Providing evidence to support different means for the different subgroups supports known-groups validity.

13.3.2.5 Ability to Detect Change (Responsiveness)

While it is important to provide evidence that the measure is assessing the concepts of interest, it is equally important within the clinical trial environment to provide evidence that scores on the PRO measure show adequate responsiveness—that is, that the PRO measure scores have the ability to detect within-subject change where change is expected. If a measure is not capable of measuring change within the time period anticipated in a clinical trial setting, the measure will not be useful for its intended purpose.

Various methods are used to evaluate responsiveness. These include correlational analyses, computation of effect sizes (standardized measures of effect or change), and reporting of standardized response means (McLeod et al. 2016). Correlational analyses can build upon the correlations computed within the construct validity evaluations. Specifically, a priori hypotheses are defined that outline the direction and magnitude of correlation coefficients for change scores based on the PRO measure and change scores for other supporting measures, where larger and positive correlation values are anticipated for similar construct pairs based on time points where change is expected (e.g., change from baseline to end of treatment using an efficacious treatment).

In addition to the correlational evaluations, effect size estimates of change are often computed to provide a standard unit to report within-group change on the new PRO measure. McLeod et al. (2016) provide an overview of the types of effect sizes that can be used when evaluating change. For example, one standard effect size of change is defined as the mean change from baseline to the end of the key evaluation period divided by the standard deviation (SD) of the baseline score. Another effect size used to evaluate responsiveness is Guyatt's responsiveness statistic, which is computed using the same numerator (mean change from baseline to the end of the key evaluation period) but the denominator is the standard deviation of the stable subgroup. Standardized response means provide a further metric to describe respon-

siveness. This statistic is often computed as the mean change in the new PRO measure score from baseline to the end of the evaluation period divided by the SD of the change score.

When reviewing responsiveness evidence based on effect sizes, it is important to understand the unit used in the denominator. It is also important to provide evidence that the change scores are different for groups that are hypothesized to be different. For example, showing that change scores are statistically different between groups identified as different by supporting measures such as a patient global impression of change provides evidence for the measure's ability to detect change.

13.3.3 Other Considerations for Psychometric Evaluation

As outlined in the preceding section, a typical psychometric evaluation focuses on cross-sectional methods for construct validity and two time points for the evaluation of test-retest reliability and responsiveness. Given the longitudinal nature of clinical trials and the need for PRO measures to measure change over time, longitudinal methods are evolving to incorporate data from more than two time points and include depictions of individual trajectories of change for multiple measures, where appropriate (Williams et al. 2015). In addition to the focus on the psychometric properties, methods like joint mixed models for repeated measures can also help researchers understand how changes in PRO scores relate to other clinical endpoints; this approach can provide valuable insight into treatment efficacy, especially for conditions where the relationships between these measures change over time (Odom et al. 2017).

The amount of missing data should be evaluated within the context of the psychometric evaluation as well as within the context of the future use of the PRO measure. There is no statistical method that can be used to handle all types of missing data. However, if a large number of item-level responses are missing within the psychometric evaluation data, the feasibility of utilizing the PRO measure in a pivotal study should be questioned. Given the relationships among the items, rules should be determined both for how to score the PRO measure when item-level responses are missing and for how much missing item-level data is "too much" and the PRO measure score should be set to missing. These rules must be justified (Chen et al. 2014).

13.3.4 Meaningful Score Change

While key endpoints in clinical trials are commonly assessed through the statistical comparison of group means, the results of these comparisons are not always easy to interpret. In addition, statistical significance alone is not sufficient to demonstrate clinically important benefit. As such, characterizing meaningful change for individual

patients and at a group level provides the ability to further evaluate, interpret, and communicate PRO results to regulators, patients, and prescribers.

The FDA guidance recommends that researchers identify the magnitude of change that is meaningful on the PRO measure at the individual level (responder definition or threshold); as such, key endpoints based on a PRO measure commonly involve comparing the proportion of patients who achieve this level of change (characterized as responders) while on an active treatment to the proportion of responders in the placebo or comparative treatment group. Other regulatory reviews may focus on determining whether the difference in benefit achieved between treatment groups (group-level means) is both statistically significant and clinically meaningful. The literature related to interpreting change on PRO measures is vast and many terms have been used to refer to this change. For individuals, the terms "responder definition," "responder threshold," and "clinically important difference" have been used; for groups, suggestions include "minimal important difference," "minimal clinically important difference," and "clinically important difference" (Coon and Cappelleri 2016; Cook et al. 2015; Fayers and Hays 2014; King 2011; Marquis et al. 2004; Revicki et al. 2007). Given the variety of methods and labels, it is essential to understand how the "level of change" was defined on a specific PRO measure prior to drawing conclusions on results.

13.3.4.1 Methods for Interpreting Individual-Level Meaningful Change

The PRO guidance (FDA 2009) outlines three methods related to establishing a responder definition or threshold for meaningful change at the patient level ("within-person meaningful change"): anchor-based, distribution-based, and cumulative distribution plots. The application of these methods has evolved and is influenced by the constructs measured and the available supporting measures within the data used to define the threshold. The reader is encouraged to consult Chap. 6 of Volume 2 (Cappelleri and Bushmakin 2018) of this series for an in-depth discussion of both traditional and emerging methods. For the purposes of this chapter, we outline one approach to defining a responder threshold intended to inform future pivotal clinical trial endpoints based on the current regulatory environment.

The primary method to estimate a responder threshold is the anchor-based method (Coon and Cappelleri 2016; FDA 2009; Guyatt et al. 2002; Cella et al. 2002; Wyrwich et al. 2013; McLeod et al. 2011). Given that the target measure is patient-reported, it is generally preferable for the primary anchor to be patient-reported; however, other types of measures with commonly accepted thresholds are also used. For example, responder definitions for PRO measures of patient functioning related to improvements in depression or weight loss may be developed by examining scores for patients with a 50% reduction in Hamilton Depression Rating Scale scores or 5% weight loss, respectively (Bobo et al. 2016; FDA 2007).

Patient Global Impression of Severity (PGIS) and Patient Global Impression of Change (PGIC) items are commonly used as anchor measures. A typical PGIS item asks patients to rate their current disease severity at the beginning of a study and at

key time points using, for instance, a 4- or 5-point response scale (e.g., "*Overall, how would you rate your psoriasis symptoms now?*" with response options "none," "mild," "moderate," "severe," or "very severe"). A typical PGIC item asks patients to rate change in their disease severity in comparison to the start of the study using a 7-point response scale that includes categories for improvement and worsening (e.g., "*Overall, compared to the start of the study, how would you rate your psoriasis symptoms now?"* with response options "much better," "moderately better," "a little better," "no change," "a little worse," "moderately worse," or "much worse"). To minimize measurement error associated with the PGIC's lengthy recall period, the PGIS is generally preferred by the FDA as the primary anchor measure. However, the PGIC and other supportive anchors (including as other established PRO measures, global items completed by clinicians or caregivers, and clinical outcomes) should be included in the analysis as appropriate to provide relevant information about meaningful change.

An anchor measure should be easier to interpret than the target PRO measure (FDA 2009) and should be evaluated for its appropriateness before it is used to estimate a responder threshold. Methods to evaluate appropriateness include descriptive statistics of the change in the PRO measure scores, correlation coefficients for change, cumulative distribution function (CDF) plots, and probability density function (PDF; also referred to as kernel density) plots for each level defined by the anchor. For an appropriate anchor, the size and direction of the mean (or median) change on the PRO measure scores should follow a predictable pattern for the anchor levels—that is, the largest positive change (improvement) in the PRO measure scores should be achieved by patients who report that they have improved the most as defined by the PGIS change and the largest negative change (decline) in PRO measure scores should be associated with patients who report that they have worsened the most as defined by the PGIS change. In addition to evidence based on the descriptive statistics pattern, appropriate anchor measures should have appreciable correlation with the target PRO measure (correlation >0.3; Revicki et al. 2008). Finally, for an appropriate anchor, the CDF and PDF plots of change in the PRO measure by the anchor categories should not overlap and, again, greater change on the anchor should align with greater change on the PRO measure scores.

After a candidate anchor measure has been evaluated and deemed appropriate, it can be applied to estimate a responder threshold. For regulatory purposes, one way that has been used to define a responder threshold is as the mean (median) change score for the subgroup reporting, for example, a single-category improvement on the PGIS between baseline and a key time point. Supportive threshold estimates are computed similarly using supportive anchors that have been evaluated and determined to be appropriate. For example, a supportive threshold based on the PGIC may be defined as the mean (median) change scores for patients reporting "moderately better" on the PGIC, one-category improvement on the clinician-rated CGI-S, or "moderately better" on the CGI-I.

In addition to the anchor-based method, it is important to include a distribution-based method when defining a responder threshold, as outlined in the FDA PRO guidance (FDA 2009). These thresholds provide additional support to the primary

anchor-based threshold by including information related to the variability within the PRO measure scores and the statistical significance of individual change. For example, a commonly computed supportive threshold is one-half standard deviation of the PRO measure scores at baseline, which corresponds to a moderate effect size (Norman et al. 2003; McLeod et al. 2011; Wyrwich et al. 2015).

Another relevant distribution-based threshold is the standard error of measurement (SEM) which is computed as $\text{SEM} = \text{SD}\sqrt{(1-r)}$, where SD is the standard deviation of the PRO measure scores at baseline and r is the test-retest (or internal consistency) reliability estimate of the PRO measure (Wyrwich et al. 1999). Because it includes reliability in its computation, the SEM-based threshold provides an estimate influenced by the measurement precision of the PRO measure. Finally, the reliable change index (RCI) is another supplemental, distribution-based method to consider. The RCI is computed as $\sqrt{(2) \times \text{SEM}}$, which provides an adjustment to the SEM and provides a z-test of the change (Hays et al. 2005).

Although not mentioned in the PRO guidance, additional supportive thresholds can be identified based on qualitative information as mentioned in Sect. 13.2. Specific questions can be asked to ascertain the amount of change on a PRO measure's individual items and sets of items that is deemed meaningful to patients in order to provide supportive evidence for responder definitions.

After the various thresholds have been estimated, it is typical to report these together, with the primary threshold based on the primary anchor informing the endpoint for the subsequent pivotal trial. The supportive threshold values should be relatively close to the primary threshold value, but it is extremely unlikely that they will have the same value. Rather, the supportive anchors provide the regulatory reviewers with an indication of the "robustness" of the proposed anchor. The regulatory reviewers may ask, for example, that one of the supportive anchor values be applied to the pivotal data as a check of the strength of the treatment's efficacy as measured by the PRO measure.

To utilize the responder definition within the pivotal trials, a statistical evaluation may be performed that compares the proportion of responders based on the PRO measure scores by treatment groups at the key time point. For this comparison, a statistically significantly larger proportion of responders within the treatment group would provide evidence for treatment benefit.

To provide additional information surrounding the responder threshold, it is also typical to provide CDFs by treatment group, with reference lines plotted for the primary and supportive threshold values. These plots facilitate the evaluation of treatment response across a range of potential responder definitions. For this application, a CDF is plotted for each treatment group and portrays the percentage of patients in each group who achieve a given change from baseline on the PRO measure. Differences between the treatment groups are depicted by the lack of overlap in the curves (the difference on the vertical axis for each PRO measure difference as represented for a score difference defined by the horizontal axis). If negative change on the PRO measure indicates improvement over time, a more efficacious treatment group's curve will be observed to the left of the placebo or comparator treatment group.

13.3.4.2 Methods for Interpreting Group-Level Meaningful Change

While the FDA has encouraged the use of methods focused on the meaningfulness of individual change, additional stakeholders (including European regulatory bodies) have encouraged methods related to group-level meaningful change. Multiple methods have been proposed for establishing this unit, including mapping PRO score differences to clinically relevant outcomes (Revicki et al. 2008; Cappelleri et al. 2013), distribution-based methods such as effect sizes, comparisons of cumulative distribution functions, mediation models, and probability of relative benefit (McLeod et al. 2016).

13.3.5 Documentation of the Psychometric Evaluation Process and Final PRO Measure

As with the development process, it is important to document all steps involved in the psychometric evaluation either separately or in a single report which covers both the development and evaluation. The documentation should include a comprehensive description of the study design used for the psychometric evaluation, participants, methodology, and results. In addition to documenting the psychometric evaluation, a formal user's manual is recommended to describe the concept(s) addressed by the PRO measure, specific information about how it should be administered, and the method for scoring (including the proposed responder threshold, as appropriate, based on the type of planned endpoint). This information will facilitate an easier application of the measure for use within a clinical trial. The manual should provide explicit instructions related to how the measure should be displayed (if on paper or electronically), including a copy of the final item wording and response category layout. In addition, any training that patients or sites will need related to administration timing or instructions should be documented in a manner that permits inclusion in a protocol appendix and regulatory briefing book. Scoring should also be described, including the handling of missing item-level responses.

In addition to summarizing all of the qualitative and quantitative evidence pertaining to the development and psychometric evaluation of the PRO measure and a copy of the PRO measure itself (complete with scoring details), regulatory submissions should include additional information pertaining to the analysis of the PRO data within the context of the pivotal trials and desired labeling claims. The inclusion of all of these items will help to facilitate regulatory review and decisions regarding the ability of the PRO measure to support product approval and labeling.

Acknowledgements We thank Lauren Nelson for helpful comments on earlier versions of this chapter. In addition, we thank Lindsey Norcross and Jason Mathes for their editorial and graphical support.

References

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills: Sage.

Baldwin, M., Spong, A., Doward, L., & Gnanasakthy, A. (2011). Patient-reported outcomes, patient-reported information: From randomized controlled trials to the social Web and beyond. *Patient, 4,* 1–7.

Bentler, P. M. (1989). *EQS structural equations program manual*. Los Angeles: BMDP Statistical Software.

Bobo, W. V., Angleró, G. C., Jenkins, G., Hall-Flavin, D. K., Weinshilboum, R., & Biernacka, J. M. (2016). Validation of the 17-item hamilton depression rating scale definition of response for adults with major depressive disorder using equipercentile linking to clinical global impression scale ratings: analysis of pharmacogenomic research network antidepressant medication pharmacogenomic study (PGRN-AMPS) data. *Human Psychopharmacology, 31,* 185–192.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park: Sage.

Cappelleri, J. C., & Bushmakin, A. G. (2018). Advancing interpretation of patient-reported outcomes. In K. Peace, D.-G. Chen, & S. Menon (Eds.), *Biopharmaceutical Applied Statistics Symposium*, Vol. 2. pp. 69–89

Cappelleri, J. C., & Spielberg, S. P. (2015). Advances in clinical outcome assessments. *Therapeutic Innovation and Regulatory Science, 49,* 780–782.

Cappelleri, J. C., Zou, K. H., Bushmakin, A. G., Alvir, J. M. J., Alemayehu, D., & Symonds, T. (2013). *Patient-reported outcomes—measurement, implementation, and interpretation*. Boca Raton, Florida: Chapman and Hall/CRC Press.

Cappelleri, J. C., Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. *Clinical Therapeutics, 36,* 648–662.

Cella, D., Bullinger, M., Scott, C., Barofsky, I., Clinical Significance Consensus Meeting Group. (2002). Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Mayo Clinic Proceedings 77*, 384–392.

Chen, W. C., McLeod, L. D., Nelson, L. M., Williams, V. S., & Fehnel, S. E. (2014). Quantitative challenges facing patient-centered outcomes research. *Expert Review Pharmacoecon Outcomes Research, 14*(3), 379–386.

Cook, K. F., Victorson, D. E., Cella, D., Schalet, B. D., & Miller, D. (2015). Creating meaningful cut-scores for Neuro-QOL measures of fatigue, physical functioning, and sleep disturbance using standard setting with patients and providers. *Quality of Life Research, 24,* 575–589.

Coon, C. D., & Cappelleri, J. C. (2016). Interpreting change in scores on patient-reported outcome instruments. *Therapeutic Innovation and Regulatory Science, 50,* 22–29.

Coons, S. J., Gwaltney, C. J., Hays, R. D., et al. (2009). Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO good research practices task force report. *Value Health, 12,* 419–429.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: A practical guide*. Cambridge: Cambridge University Press.

DeMuro, C. D., Lewis, S. A., DiBenedetti, D. B., Price, M. A., & Fehnel, S. E. (2012). Successful implementation of cognitive interviews in special populations. *Expert Review Pharmacoecon Outcomes Research, 12*(2), 181–187.

Deyo, R. A., Diehr, P., & Patrick, D. L. (1991). Reproducibility and responsiveness of health status measures: Statistics and strategies for evaluation. *Controlled Clinical Trial, 12,* 142S–158S.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16,* 5–18.

European Medicines Agency (EMA). (2005). *Reflection paper on the regulatory guidance for the use of health related quality of life (HRQL) measures in the evaluation of medicinal products*. London: European Medicines Agency.

Fayers, P. M., & Hays, R. D. (Eds.). (2005). *Assessing quality of life in clinical trials: Methods and practice*. Oxford: Oxford University Press.

Fayers, P. M., & Hays, D. R. (2014). Don't middle your MIDs: regression to the mean shrinks estimates of minimally important differences. *Quality of Life Research, 23,* 1–4.

Fayers, P. M., & Machin, D. (2016). *Quality of life: The assessment, analysis and reporting of patient-reported outcomes* (3rd ed.). Chichester: Wiley.

Food and Drug Administration (FDA). (2007). Guidance for industry. Developing products for weight management. https://www.fda.gov/downloads/Drugs/Guidances/ucm071612.pdf. Accessed June, 01 2017.

Food and Drug Administration (FDA). (2009). Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf. Accessed January, 3 2017.

Food and Drug Administration (FDA). (2013a). Roadmap to patient-focused outcome measurement in clinical trials. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm. Accessed January 5, 2017.

Food and Drug Administration (FDA). (2013b). Center for Drug Evaluation and Research. Drug Development Tool Number: COA DDT 003 Study Endpoints and Labeling Development (SEAL) Review. SEALD Tracking Number: 2013–055. http://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/UCM386244.pdf. Accessed January 28, 2017.

Food and Drug Administration (FDA). (2016). Clinical outcome assessment compendium. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DevelopmentResources/ucm459231.htm. Accessed January, 8 2017.

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D.; Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health 10*, S94–S105.

Gnanasakthy, A., Mordin, M., Clark, M., et al. (2012). A review of patient-reported outcome labels in the United States: 2006–2010. *Value Health, 15*(3), 437–442.

Gnanasakthy, A., Mordin, M., Evans, E., Doward, L., & DeMuro, C. (2017). A review of patient-reported outcome labeling in the United States (2011–2015). *Value Health, 20*(3), 420–429. https://doi.org/10.1016/j.jval.2016.10.006.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale: Lawrence Erlbaum.

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., Norman, G. R.; Clinical Significance Consensus Meeting Group. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clin Proceedings 77*, 371–383.

Hays, R. D., Brodsky, M., Johnston, M. F., Spritzer, K. L., & Hui, K. (2005). Evaluating the statistical significance of health-related quality of life change in individual patients. *Evaluation and the Health Professions, 28,* 160–171.

Hays, R. D., Revicki, D. (2005). Reliability and validity (including responsiveness). In P. M. Fayers, R. D. Hays (Eds.) Assessing quality of life in clinical trials: methods and practice. Oxford: Oxford University Press, pp. 25–39.

King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. *Expert Review Pharmacoecon Outcomes Research, 11,* 171–184.

Marquis, P., Chassany, O., & Abetz, L. (2004). A comprehensive strategy for the interpretation of quality-of-life data based on existing methods. *Value Health, 7,* 93–104.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

McLeod, L. D., Cappelleri, J. C., & Hays, R. D. (2016). Best (but oft-forgotten) practices: expressing and interpreting associations and effect sizes in clinical outcome assessments. *The American*

Journal of Clinical Nutrition 103(3), 685–693 (with erratum in *The American Journal of Clinical Nutrition* 2017;105:241).

McLeod, L. D., Coon, C. D., Martin, S. A., Fehnel, S. E., & Hays, R. D. (2011). Interpreting patient-reported outcome results: US FDA guidance and emerging methods. *Expert Review Pharmacoecon Outcomes Research, 11,* 163–169.

Messick, S. (1989). Validity. *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality-of-life: The remarkable universality of half a standard deviation. *Medical Care, 4,* 582–592.

Norquist, J. M., Girman, C., Fehnel, S., DeMuro-Mercon, C., & Santanello, N. (2012). Choice of recall period for patient-reported outcome (PRO) measures: Criteria for consideration. *Quality of Life Research, 21*(6), 1013–1020.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Odom, D., McLeod, L., Sherif, B., Nelson, L., McSorley, D. (Under review). Longitudinal modeling approaches to assess the association between changes in a patient-reported outcome and a clinical endpoint.

Odom, D., McLeod, L., Sherif, B., Nelson, L., McSorley, D. (2017). Longitudinal modeling approaches to assess the association between changes in 2 clinical outcome assessments. *Ther Innov Regul Sci*. 2017 Sep 26.

Patrick, D. L., Burke, L. B., Gwaltney, C. J., et al. (2011a). Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1–eliciting concepts for a new PRO instrument. *Value Health, 14,* 967–977.

Patrick, D. L., Burke, L. B., Gwaltney, C. J., et al. (2011b). Content validity–establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 2—assessing respondent understanding. *Value Health, 14,* 978–988.

Reeve, B. B., Wyrwich, K. W., Wu, A. W., et al. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Quality of Life Research, 22,* 1889–1905.

Revicki, D. A., Erickson, P. A., Sloan, J. A., et al; Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group. (2007). Interpreting and reporting results based on patient-reported outcomes. *Value Health 10*, S116–24.

Revicki, D., Hays, R., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61,* 102–109.

Rothman, M., Gnanasakthy, A., Wicks, P., & Papadopoulos, E. J. (2015). Can we use social media to support content validity of patient-reported outcome instruments in medical product development? *Value Health, 18,* 1–4.

Schuck, P. (2004). Assessing reproducibility for interval data in health-related quality of life questionnaires: Which coefficient should be used? *Quality of Life Research, 13,* 571–586.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). New York: Oxford University Press.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design*. San Francisco: Jossey-Bass.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Walton, M. K., Powers, J. H., Hobart, J., et al. (2015). Clinical outcome assessments: Conceptual foundation. Report of the ISPOR clinical outcomes assessment – Emerging good practices for

outcomes research task force. *Value Health 18*, 741–752. https://doi.org/10.1016/j.jval.2015.08.006.

Wild, D., Grove, A., Martin, M., et al. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value Health, 8,* 94–104.

Williams, V., McLeod, L., & Nelson, L. (2015). Advances in the evaluation of longitudinal construct validity of clinical outcome assessments. *Therapeutic Innovation and Regulatory Science, 49,* 805–812.

Willis, G. B. (2005). *Cognitive interviewing: a tool for improving questionnaire design*. Thousand Oaks: Sage.

Willis, G. B. (2015). Analysis of the cognitive interview in questionnaire design. understanding qualitative research. New York: Oxford University Press.

Wyrwich, K. W., Norquist, J. M., Lenderking, W. R., Acaster, S.; Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL). (2013). Methods for interpreting change over time in patient-reported outcome measures. *Quality of Life Research 22*, 475–483.

Wyrwich, K. W., Krishnan, S., Poon, J. L., et al. (2015). Interpreting important health-related quality of life change using the Haem-A-QoL. *Haemophilia, 21,* 578–584. https://doi.org/10.1111/hae.12642.

Wyrwich, K. W., Tierney, W. M., Wolinsky, F.D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol, 52*:861–873.

Chapter 14
Interim Analyses: Design and Analysis Considerations for Survival Trials When Hazards May Be Nonproportional

Edward Lakatos

14.1 Introduction

In the decade following the introduction of the group-sequential concept by Pocock (1977), many publications emerged investigating and furthering this concept. In the late 1980s publications by Bauer (1989) ("Multistage Testing with Adaptive Design"), Wittes and Brittain (1990) ("The role of internal pilot studies in increasing the efficiency of clinical trials"), as well as Gould and Shih (1992), Gould (1992) and Shih (1993) ushered in the era of adaptive designs—statistical research in this area flourished for the next two decades. Group-sequential methods became viewed as part of the broader category of adaptive methods.

Non-proportional hazards was recognized as an important factor for the design of survival trials as far back as 1968 (Halperin et al., for sample size methodology for a comparison of proportions for two types of treatment lags). Lakatos (1986, 1988) further contributed to sample size methodology providing methodology for the logrank statistic for unrestricted non-proportional hazards alternatives.

For survival analysis, methodology for weighting the logrank statistic to deal with NPH was proposed by Tarone and Ware (1977), Harrington and Fleming (1982), Zucker and Lakatos (1990), Self (1991) and Yang and Prentice (2010).

In contrast, except for publications by Lakatos (2002, 2015, 2016), there has been relatively little mention of non-proportional hazards in the adaptive arena.

The focus of this chapter will be on survival trials, especially when the proportional hazards assumption is suspect. Nonproportional hazards can occur in many areas of medicine.

Earlier publications by the author focusing on trials with non-proportional hazards, dealt with developing statistical methods, as well as answering such questions as "why it is important to recognize and address non-proportionality", and how meth-

E. Lakatos (✉)
BiostatHaven, Inc., Croton-on-Hudson, USA
e-mail: elakatos@msn.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_14

ods work. This chapter focuses on which methods to use, and when and how to apply such methods.

The chapter begins (Sect. 14.2) with a published oncology trial, designed using exponential assumptions, that turned out to have strongly non-proportional hazards. One of the interesting features of that non-proportionality is that it possesses both a delayed treatment effect, and a treatment effect that disappears: the survival curves coincide for periods both at the beginning and before the end of the trial. When separated survival curves come together, the hazard ratio is in the wrong direction, and the logrank statistic is negatively affected. This does not necessarily indicate an untoward effect of treatment, but could simply reflect that the treatment successfully delayed an event in the sickest patients, resulting in an imbalance, with more sick patients remaining at risk in the experimental arm. Depending on how the trial is designed (for example, the lengths of the recruitment and total trial), the survival experience can put more weight on the earlier or later portions of the survival curves. For example, if a small number of patients is enrolled, and one simply waits until enough events occur, this could end up with most of the patients residing for long periods in the no treatment effect zone after the curves come back together. This will have a detrimental effect on the logrank statistic. The situation of the trial discussed in the next section is more complex. That section shows, in a series of steps, how a "designer" group sequential boundary can enhance the trial's chances of ending successfully early, with a far smaller sample size compared with the original.

Section 14.3 discusses spending functions for designer boundaries.

The concept of a "Binding Boundary" emerged in the early 2000s. This term was used to challenge the legitimacy of a large number of designs that had been the mainstay of much of the group-sequential literature. The concept does not appear to have been discussed in the literature. In addition, the basic problem is not restricted group-sequential methods. Section 14.4 discusses the concept, and points to a history and recurrence of designs that share the same basic flaw of the binding boundary.

If one wishes to stop a trial early for futility, there is a choice between futility boundaries and conditional power. Section 14.5 discusses pros and cons.

When hazards are non-proportional, power can often be improved through the use of a weighted logrank statistic. Section 14.6 discusses optimal weighting for a delayed treatment effect.

14.2 Designer Boundaries

In this section, given a prior similar trial, the design of a new trial is discussed. The survival curves from the prior trial are quite non-proportional, leading to a variety of design considerations. A non-standard group-sequential boundary is constructed to address those considerations. A stepwise approach leading to that boundary is now described.

In designing the International Collaboration on Ovarian Neoplasms trial (ICON 7), Perren et al. (2011) assumed exponential hazards with a control group median

survival of 18 months [Progression-Free Survival (PFS)] and 43 months [Overall Survival(OS)], and an increase due to therapy to 23 months (hazard ratio 0.78) (PFS), and to 53 months (hazard ratio 0.81) (OS). For 90% power at a 2-sided 0.05 significance level, they calculated 684 PFS (715 OS) events would be needed. With 1520 women randomly assigned to treatment over a period of 2 years, the required 684 PFS events were expected to occur 3 years after study start.

The PFS analysis took place after 759 PFS events occurred; PFS was significant, while OS was not close.

Consistent with a visual inspection of the KM curves in Fig. 14.1, the results paper stated that a test showed that the hazards were not proportional. We note here that, in contrast to the proportional hazards model, the logrank test does not require proportional hazards. Although the ordinary logrank test is not optimal when proportional hazards is violated, for a given non-proportions hazards alternative, an optimal weighting for the logrank exists.

The ICON 7 trial (Perren et al. 2011) was designed to evaluate both PFS and OS. With the PFS KM curves now available, I show how the design of a new PFS trial could be approached. Analysis will use the logrank statistic (although a case for the use of a weighted logrank statistic could be made, the simpler standard logrank statistic will be used).

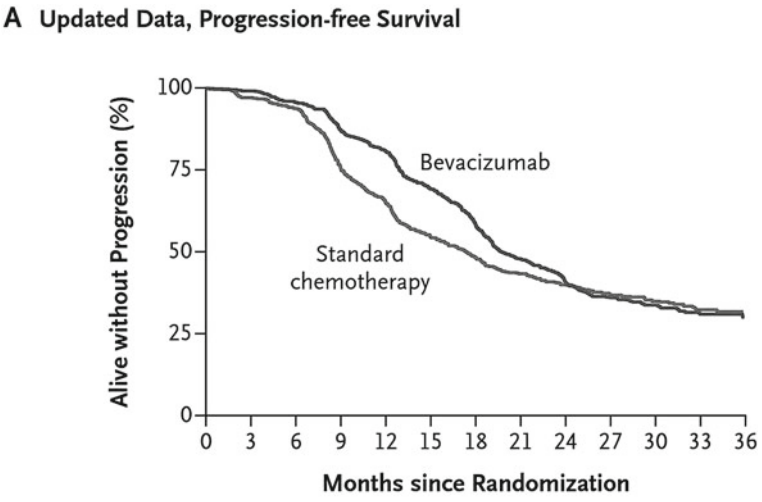


Fig. 14.1 Kaplan-Meier survival curves from Perren et al. (2011)

Table 14.1 presents time-dependent failure rates captured from the PFS KM curves of Fig. 14.1. These time-dependent rates were extracted from the published survival curves using a method developed by Lakatos (US Patent US20100063741 A1)

Table 14.2 presents sample size calculations for a fixed-sample design (no interim analyses) based on the time-dependent failure rates extracted from Fig. 14.1 as presented in Table 14.1. These sample size calculations assume 90% power, and a significance level of 0.05, 2-sided.

From the Perrin publication, the original sample size calculations, based on exponential assumptions, for a trial lasting 36 months with a recruitment period of 24 months, led to the 684 events required for the PFS endpoint. The same assumptions, except for using the time dependent rates of Table 14.1, leads to 877 required events (Lakatos 1988 as implemented in STOPP®). Under the exponential assumption, the number of events is independent of the recruitment and trial lengths, i.e., all of the entries in the "events" panel of Table 14.3 would be 877.

But the number of events in that upper panel range from 122 to 2063. For a 12-month trial, the 122 events is paired with a sample size of 1226 in the lower panel.

Table 14.1 Annual failure rates for the designated periods, derived from Fig. 14.1

Month	Standard	Bevacizumab	Hazard ratio
0–6	0.1177	0.0815	0.6827
6–12	0.5037	0.2873	0.4829
12–18	0.4638	0.4669	1.0096
18–24	0.3108	0.5089	1.9129
24–30	0.2387	[a]0.2387	1.4969
30–36	0.1962	[a]0.1962	0.7066

[a]The survival curves appear to be identical during months 24–36

Table 14.2 Sample size calculations (PFS) based on time-dependent failure rates in Table 14.1 (unweighted logrank)

Recruitment length (months)	Trial length (months)			
	12	18	24	36
Events				
12	122	164	395	2063
18		168	303	1457
24			303	877
36				675
Sample size				
12	1226	597	867	3213
18		872	839	2402
24			1089	1597
36				1691

Table 14.3 Designing a boundary for PFS based on the time-dependent survival curves of Fig. 14.1 from Perron et al. Step 1

Markov projections[a]						Group-sequential calculations		
Month	InfoFrac	Events			Patients	Upper boundary		Power
		Exper	Cntrl	Total	Recruit	Z	Prob	
9.4	0.2	13.1	19.6	32.7	482	4.875	5.39E−07	0.02
12.3	0.4	26.2	39.2	65.4	598	3.3571	0.0004	14.5
14.5	0.6	39.3	58.8	98.1	598	2.6803	0.0038	54.4
16.2	0.8	52.4	78.5	131.0	598	2.2898	0.0122	78.9
18.0	1	65.6	98.2	163.8	598	2.0310	0.0250	90.0

[a]Recruitment: 12 months, total trial: 18

While less than the 1520 of the actual trial, a much smaller sample size paired with a larger number of events is possible with an 18 month trial in which recruitment is complete in 12 months: sample size 597, with 164 events. This is a dramatic reduction from the original sample size calculations.

The reduction in the required number of events with a shorter trial runs counter to usual expectations. Under the clearly violated exponential model, the number of events is independent of the trial or recruitment length. The smaller number of events for shorter trials can be explained by examining the survival curves. In particular, they initially separate, but then start coming together at about 12–15 months. From the discussion of the logrank statistic in Sect. 14.6, it is easy to see that events which occur during periods in which there is no treatment effect, or the effect is in the wrong direction, decrease the significance of the logrank statistic. The number of events that occur in such periods generally increases with trial length for the ICON 7 trial (Fig. 14.1), which explains why the required number of events increases as trial length increases.

When the recruitment period is shortened, on average, patients move to the latter part of the trial more quickly, and reside there longer. This, in turn, results in more events occurring when there is no treatment effect, or the treatment effect is in the wrong direction. As above, this decreases the significance of the logrank statistic.

In designing the new PFS trial, due to variability, caution should be used to avoid over-reliance on the survival curves from Perrin presented in Fig. 14.1. However, it is reasonable to expect the underlying PFS curves for the new trial will be much more likely to resemble the non-exponential shapes of Fig. 14.1 than being exponential. And the curves in Fig. 14.1 are based on a large number of patients (1520) and events (759).

One of the design challenges is as follows. Referring to Table 14.2, if the 12/18 months (recruitment/trial length) design is implemented, and recruitment takes 24 months, then a larger number of events would have been appropriate. One way to address this problem of uncertainty is through Designer Group-Sequential Boundaries.

The discussion now turns to developing such a "Designer Boundary" to improve the design. This will be achieved through judiciously varying the lengths trial and recruitment periods, and modifying the spending function to take advantage of periods of strong treatment effect.

A sequence of tables, each providing projected operating characteristics of a group-sequential design is now presented. Each such table represents a modification of the design of the preceding tables, with the intent of addressing some issue. It shows how one could achieve objectives by modifying the group-sequential procedure in atypical ways.

Table 14.3 presents results of calculations based on the time-dependent failure rates of Table 14.1. The Group-Sequential module of STOPP® was used throughout. Assumptions are the usual two-sided 0.05 significance level, 90% for the fixed-sample design power, and additionally, 5 looks to occur at equally spaced increments of information fractions. The design assumes 12 months uniform recruitment, 18 months total trial length.

The Markov model is used to project all of the columns labeled "Markov Projections". The software allows the user to specify the interim plan in terms of either (1) desired months, or (2) desired information fractions, and the Markov portion of the program will calculate the remaining columns under "Markov Projections" (Lakatos 2002).

The group-sequential plan can be specified through a variety of possible spending functions or a custom spending function (discussed below); the cumulative alphas appear in the column labeled "Prob" (cumulative probability of a Type I error). When the spending function is formula-based, that formula applied, to the information fractions in column 2, determines the values in the "Prob" column. For a custom spending function, any cumulative alpha values can be manually entered, provided they satisfy the criteria for a spending function (defined in Sect. 14.3). The Z-values for the boundary are calculated from the information fractions and corresponding cumulative alphas.

With uniform information fractions, the 4th interim will occur only 1.8 months before the final, which is too close to be of practical value. Table 14.4 explores the option of performing the interim in relatively uniform increments of calendar time.

Table 14.5 explores the implications of the recruitment taking 18 months, here using uniform information fractions.

These sample sizes and numbers of events are far less than the published prior trial. If the survival curves of the new trial deviate from those of Fig. 14.1, a longer trial may be needed. Table 14.6 explores the operating characteristics of a trial lasting 24 months, with recruitment complete in 18.

The required number of events for this trial, 287, has increased substantially over the 164 of Tables 14.3 and 14.4. Table 14.6 trial is likely (83.9% power) to end at the 0.8 information fraction look, at 230 events. But it is only 2.5 months short of the maximum 24 months for this trial, which again is not practical.

Table 14.7 reduces the number of interims to 4, replacing the interims at 0.6 and 0.8 with a single interim at information fraction 0.7. The timing of the next-to-last

Table 14.4 Designing a boundary for PFS based on the time-dependent survival curves of Fig. 14.1 from Perron et al. Step 2

Markov projections[a]						Group-sequential calculations		
Month	InfoFrac	Events			Patients	Upper boundary		Power
		Exper	Cntrl	Total	Recruit	Z	Prob	
4	0.02	1.55	2.32	3.87	199.02	6	0	0
7	0.08	4.96	7.42	12.38	348.29	6	2.22E−15	0
11	0.29	18.77	28.09	46.86	547.32	4.0010	3.15E−05	2.0
14	0.55	35.83	53.65	89.48	597.07	2.8074	0.0025	45.0
18	1	65.59	98.2	163.79	597.07	1.9740	0.0250	90.0

[a]Recruitment: 12 months, total trial: 18

Table 14.5 Designing a boundary for PFS based on the time-dependent survival curves of Fig. 14.1 from Perron et al. Step 3

Markov projections[a]						Group-sequential calculations		
Month	InfoFrac	Events			Patients	Upper boundary		Power
		Exper	Cntrl	Total	Recruit	Z	Prob	SS = 875
9.9	0.2	13.4	20.0	33.4	480	4.875	5.39E−07	0
12.6	0.4	26.8	40.1	66.9	614	3.3571	0.0004	15.2
14.7	0.6	40.3	60.2	100.5	715	2.6803	0.0038	54.6
16.4	0.8	53.7	80.4	134.1	800	2.2898	0.0122	78.9
18.0	1	67.2	100.5	167.7	875	2.0310	0.0250	90

[a]Recruitment: 18 months, total trial: 18

Table 14.6 Designing a boundary for PFS based on the time-dependent survival curves of Fig. 14.1 from Perron et al. Step 4

Markov projections[a]						Group-sequential calculations		
Month	InfoFrac	Events			Patients	Upper boundary		Power
		Exper	Cntrl	Total	Recruit	Z	Prob	SS = 796
12.3	0.2	25.1	32.2	57.3	545	4.875	5.39E−07	0.26
16.1	0.4	50.2	64.5	114.7	709	3.3571	0.0004	29.8
18.9	0.6	75.4	96.8	172.2	795	2.6803	0.0038	68.1
21.5	0.8	100.6	129.2	229.8	795	2.2898	0.0122	83.9
24.0	1	125.8	287.3	287.3	795	2.0310	0.0250	90.0

[a]Recruitment: 18 months, total trial: 24

Table 14.7 Designing a boundary for PFS based on the time-dependent survival curves of Fig. 14.1 from Perron et al. Step 5

Markov projections[a]						Group-sequential calculations		
Month	InfoFrac	Events			Patients	Upper boundary		Power
		Exper	Cntrl	Total	Recruit	Z	Prob	
12.3	0.2	25.3	32.5	57.9	550	4.875	5.39E−07	0.27
16.0	0.4	50.8	65.2	115.9	716	3.3570	0.00039	30.29
20.2	0.7	88.9	114.2	203.1	803	2.445	0.00738	78.16
24.0	1	127.1	163.2	290.3	803	2.0005	0.025	90.0

[a]Recruitment: 18 months, total trial: 24

Table 14.8 Designing a boundary for PFS based on the time-dependent survival curves of Fig. 14.1 from Perron et al. Step 6

Markov projections[a]						Group-sequential calculations		
Month	InfoFrac	Events			Patients	Upper boundary		Power
		Exper	Cntrl	Total	Recruit	Z	Prob	
12.	0.2	25.4	32.6	57.9	551	4.875	5.39E−07	0.25
16.1	0.4	50.8	65.2	116.1	717	3.357	0.000394	29.2
20.3	0.7	89.0	114.3	203.4	804	2.172	0.015	84.5
24.0	1	127.3	163.4	290.7	804	2.111	0.025	90.0

[a]Recruitment: 18 months, total trial: 24

interim is better, being nearly 4 months from the final; but the power at that interim is only 78% compared to 83.9 of Table 14.6.

Table 14.8 increases this power by allocating more alpha to the 3rd look: the cumulative alpha is increased from 0.00738 to 0.015. The probability of stopping by the 3rd look is now 84.5% at 203 events.

One further adjustment would be to replace the 0.7 information fraction look by one at 0.6, but still allocating 0.015 cumulative alpha to the 3rd look.

The power of the 3rd look in Table 14.9 is relatively close to the corresponding power in Table 14.8 (83.1% vs. 84.5%), but the interim occurs substantially earlier, both in months and number of events. In fact, there is now 83% power of stopping for significance after 170 events, very close to the designs of Tables 14.3 and 14.4. The advantage of Table 14.9 is that if the trial does not stop at the 3rd look, it will naturally go on to 284 events. This provides some protection against the uncertainty inherent in the survival curves observed in ICON7, and over-reliance on those observed survival curves.

Designer Boundaries Summary. If the methods employed for the original sample size were used, the resulting design would require 1520 patients and 684 PFS events. Fixed sample calculations based on the unusually shaped survival curves observed in Perrin et al. show that the sample size requirements can vary dramatically when

Table 14.9 Designing a boundary for PFS based on the time-dependent survival curves of Fig. 14.1 from Perron et al. Step 7

Markov projections[a]						Group-sequential calculations		
Month	InfoFrac	Events			Patients	Upper boundary		Power
		Exper	Cntrl	Total	Recruit	Z	Prob	
12.3	0.2	24.7	31.8	56.5	537	4.875	5.39E−07	0.25
16.1	0.4	49.5	63.6	113.1	699	3.357	0.000394	29.1
18.9	0.6	74.4	95.5	169.9	784	2.172	0.015	83.1
24.0	1	124.1	159.3	283.4	784	2.111	0.025	90.0

[a]Recruitment: 18 months, total trial: 24

the length of the total trial and recruitment periods are varied: 1597 patients and 877 events for a 36 month trial with recruitment planned for 24 months. However, if the recruitment can be completed in 12 months for an 18 month trial, the requirement drops to 597 patients and 164 events. If recruitment takes 18 months, then the number of patients needed increases to 875 with 168 events. If recruitment takes longer than 18 months, the required number of patients and events can be even larger.

Some of the uncertainty of how long the actual recruitment will take can be addressed using a Designer Group-Sequential Boundary. With the group-sequential design in Table 14.9, if the trial does take the full 18 months to complete recruitment of the 784 patients, and the trial goes to the maximum 24 months to accrue 283 events, then the power is 90%. And even if recruitment takes 18 months, there is 83% power at the 60% interim if it takes 18 months to accrue 170 events. So with the design of Table 14.9, which is much smaller (784 vs. 1520 patients original; a maximum of 283 events vs. 684 events original), and shorter (24 months maximum vs. 36 months original), the power is about 83% to end by 170 events depending on the length of recruitment.

The "custom" or "designer" boundary developed in Table 14.9 is presented in Table 14.10 and Figs. 14.2 and 14.3. Note that the O'Brien-Fleming boundary provides less (cumulative) power at the final analysis as compared to the designer boundary in spite of the fact that that last critical value for the Designer ($Z = 2.15$) is higher than for the O'Brien-Fleming ($Z = 1.98$). The reason is that the Designer boundary takes better advantage of the 60% interim analysis, by allocating more alpha to the time when the treatment effect is stronger. This is consistent with an example given by Lakatos (2015, Example 3, p. 149) in which the fixed-sample design requires a larger sample size than the group-sequential.

For comparison, the O'Brien-Fleming boundary for the designated information fractions is displayed in Table 14.10 and Fig. 14.2. Figure 14.3 presents a custom spending function giving rise to this custom boundary; this is discussed in the next section.

Table 14.10 A designer boundary for PFS based on the time-dependent survival curves of Fig. 14.1 from Perron et al.

Interim	1	2	3	4
Information fraction[a]	0.2	0.4	0.6	1
Cumulative alpha spent	5.39E−07	0.000394	0.015	0.025
Incremental alpha spent	5.39E−07	0.0000394	0.0146	0.01039
Z-value designer boundary	4.875	3.357	2.172	2.15134
Z-value O'Brien-Fleming[b]	4.875	3.357	2.68028	1.98139
Power designer boundary[c]	0.25	29	83	90
Power O'Brien-Fleming[c]	0	0.12	57	88
Events designer boundary	57	113	170	283
Events O'Brien-Fleming	7	52	151	283

[a]Note that the information fractions here are not uniformly spaced
[b]O'Brien-Fleming provided for reference
[c]Power based on designer boundary sample size, with 283 events, 784 patients

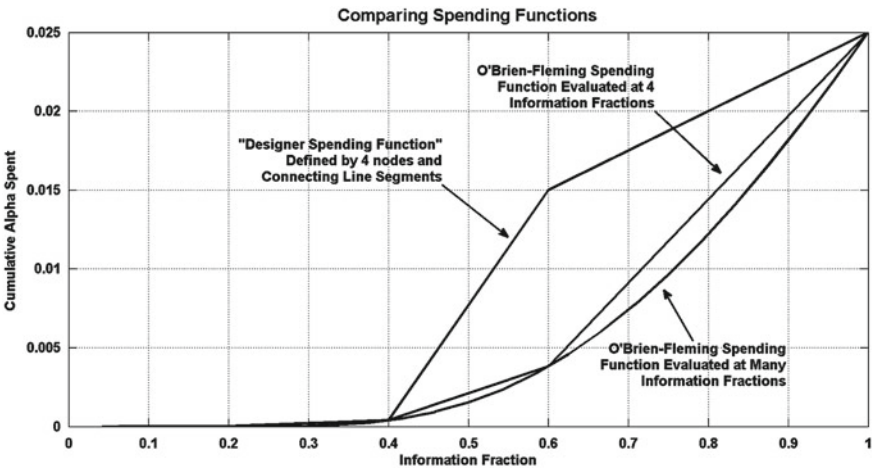


Fig. 14.2 Designer (custom) spending function compared with the O'Brien-Fleming

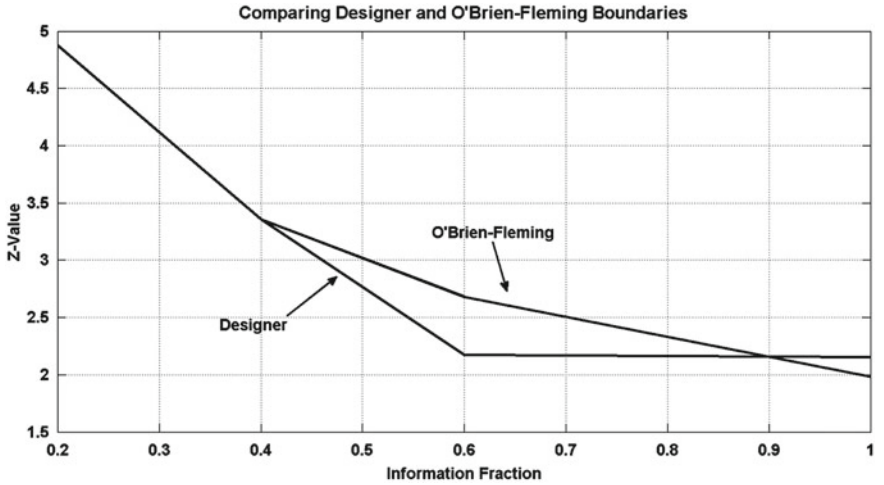


Fig. 14.3 Designer boundary compared with O'Brien-Fleming

14.3 Spending Functions for Designer Boundaries

In the publications that introduced the Pocock (1977) and O'Brien-Fleming (1979) boundaries, the procedures were based on interims taking place at equal increments of patients (e.g., 200, 400, 600, 800, 1000 patients). In addition, immediate response was implicitly assumed—that is, as soon as the patient was randomized, the response to treatment could be evaluated. At the time, an urgent need for group-sequential methods was for survival trials, for which efficacy for a given patient emerged slowly after randomization. For survival trials, the number of events, rather than patients, provides the information (Lan and Zucker 1993), and hence the timing of interim analyses. With DSMB meetings typically scheduled in equal increments of calendar time, for example—every six months, there was a disconnect between the occurrence of meetings and the accumulation of information. Both the Pocock and O'Brien-Fleming procedures were explicitly posited in equal increments of patients, with no modifications provided or discussed to accommodate deviations from those increments.

To address this problem, Lan and DeMets (1983) introduced the spending function, which generalized the group-sequential procedure to allow interims at unequal increments of information. The spending function can be viewed as a way to allocate portions of the alpha to the interim analyses, similar to the way the Bonferroni procedure can be used to allocate alpha to simultaneous tests of multiple hypotheses. An important feature of the spending function is that the increments did not have to be prespecified. So, if an efficacy analysis was to be performed at some DSMB meeting scheduled in calendar time, the number of events available at that interim could be used to allocate the alpha. The spending function must be prespecified so that the allocation of alpha is determined by the spending function and the fraction of

Table 14.11 O'Brien-Fleming boundary calculated using the spending function

Patients	200	400	600	800	1000
Information fractions	0.2	0.4	0.6	0.8	1.0
Bndry Z-values OBF Spend Fcn	4.875	3.3569	2.6803	2.2898	2.0301

information available, and cannot be manipulated by user at the time of the interim, as that could inflate the Type I error. Over time, interims for survival trials became event-driven. Still, due to the logistics of data collection and analysis, as well as the scheduling of meetings, the spending function plays an essential role in allocating alpha for the data available at the time of a DSMB meeting.

For each increment in the number of patients, the Pocock and O'Brien-Fleming boundaries provide a Z-value for testing at that interim. The Lan-DeMets spending functions corresponding to those boundaries are designed to give the same Z-values as the original procedures if the interims occur exactly at those predefined increments of patients. Because the O'Brien-Fleming boundary is an implementation of the horizontal Brownian motion boundary, the formula is well-known, and the spending function provides values that exactly match those in the O'Brien-Fleming publication. For the Pocock procedure, which specifies equal Z-values at the interims, there is no known corresponding theoretical formula—the formula is ad hoc, and does not provide exactly equal Z-values at the interims. This is immaterial, as long as one adheres to the prespecified spending function.

After the introduction of the spending function, a number of additional spending functions were proposed in the literature [for example, Kim and DeMets (1987), Hwang, Shih and DeCani (1990)]. The designer boundaries discussed in the previous section are not likely to fit into any of these previously-defined spending functions. Designer boundaries are designed to meet specific objectives and conditions that do not fit nicely within the confines of predefined spending functions.

But spending functions are easily constructed for designer boundaries. First, a spending function is defined as a monotone increasing function of the information $\alpha(x)$ that provides cumulative alpha to be spent based on the currently available fraction of information x. When $x = 0, \alpha(0) = 0$, and $\alpha(1) =$ the alpha level (typically 0.025, 1-sided). For a spending function to be consistent with a boundary such as was defined in the original Pocock or O'Brien-Fleming publications, that spending function must go through each of the coordinate pairs identified for that boundary.

For example, for a 5-interim group-sequential boundary (the 5th "interim" being the final analysis), the five Z-values for the O'Brien-Fleming boundary can be calculated using the O'Brien-Fleming spending function and the five information fractions in the second row of Table 14.11. This matches closely with the Z-values obtained using the numbers of patients in the first row and the approach presented in the original O'Brien-Fleming publication.

The Designer Boundary developed in Tables 14.3, 14.4, 14.5, 14.6, 14.7, 14.8 and 14.9 is presented in Table 14.10. The O'Brien-Fleming boundary at those information fractions is provided in Table 14.10 for comparison.

A spending function that is consistent with the Designer Boundary with 4 prespecified interims must go through the 4 coordinate pairs {(0.2, 0.000), (0.4, 0.000), (0.6, 0.015), (1.0, 0.025)} at the designated information fractions specified in Table 14.10; it can be any monotone increasing function that goes through those coordinate pairs. It is easiest to simply use straight line segments to connect the coordinate pairs. Such a function satisfies all of the criteria for a spending function.

The fact that the Z-values of the designer boundary do not exactly fit one of the published functional forms of a spending function is not an issue. The spending function for the Pocock boundary given by Lan-DeMets (1983) does not lead to exactly equal Z-values as was the specification in Pocock's original paper. If interim analyses take place at unequal information fractions, the Z-values arising from the spending function can be quite unequal. The continuity of the spending function is of great importance. If the Z-value is near the boundary at some given interim, it may seem tempting to perform another interim in rapid succession, even though this data-driven modification is not an allowable feature of the spending function. Because of the continuity of the spending function, a small increment in time will result in a correspondingly small increment in alpha, resulting in a larger Z-boundary value. This feature of the spending function provides strong protection from "cheating", even if the interims are data-driven (Proschan et al 1992).

14.4 Binding Boundaries

During the years approximately 2000–2010, the term "binding boundary" was commonly used to refer to a type of group-sequential procedure that would inflate the Type I error, and consequently was not considered valid. There has not been much, if any attention to this concept in the literature. The term "binding boundary" was introduced to describe the situation in which the validity of the Type 1 error requires strict adherence to the crossing of a lower boundary. A more detailed explanation follows.

The situation arises when the calculation of an efficacy boundary takes into account possible crossings of a futility boundary. In this case, the critical values defining the efficacy boundary can be relaxed while still maintaining the overall Type I error. The more aggressive the futility boundary, the more the efficacy critical values can be relaxed. The concept is presented diagrammatically.

Since focus will be on the Type I error, the null hypothesis will be assumed for the remainder of this section.

Begin with a fixed-sample trial (Fig. 14.4) for which the only analysis will occur at the single time of the designated end of trial. If trials are performed and analyzed a large number of times, and the Z-value computed for each trial, then some of those trials will have Z-values exceeding 1.96 (Trial B, for example), and some less than

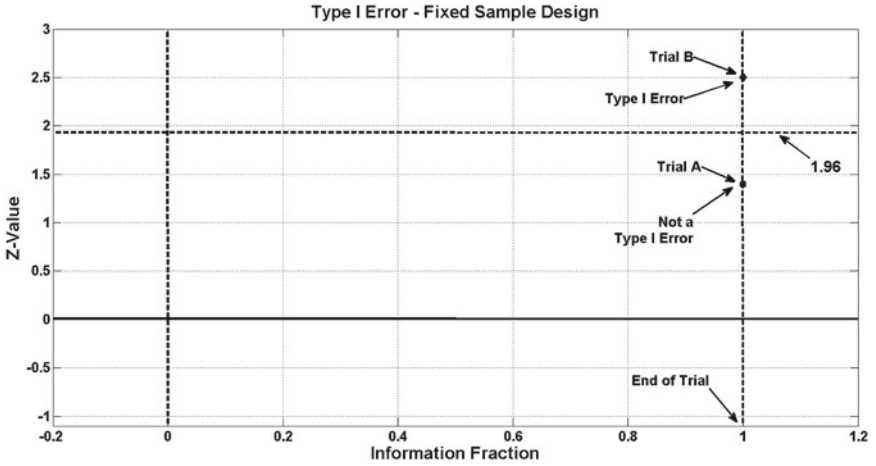


Fig. 14.4 Binding boundaries Example Part 1: fixed-sample design

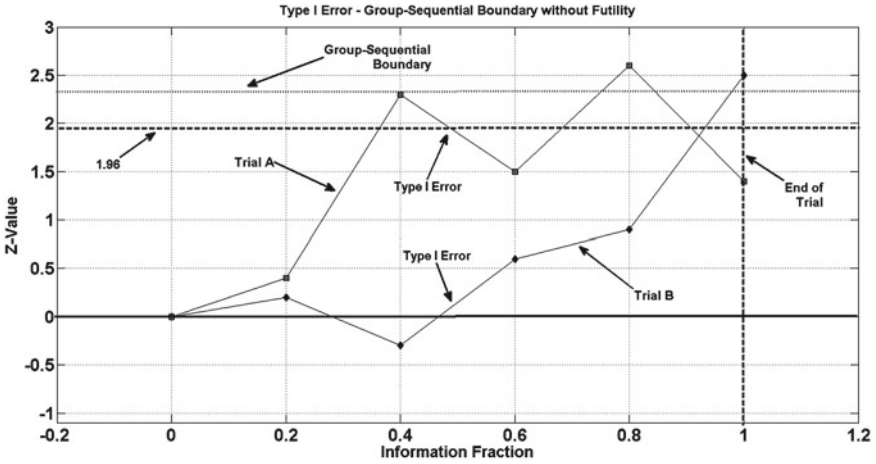


Fig. 14.5 Binding boundaries Example Part 2: A group-sequential design with no futility boundary

1.96 (Trial A, for example). Since the Z-value for Trial B exceeds 1.96, Trial B results in a Type I error, while Trial A does not. The proportion of trials exceeding 1.96 will be 0.025.

Now consider the same large number of trials, but designed as group-sequential, for which 5 interims (including the final) are planned to occur at information fractions {0.2, 0.4, 0.6, 0.8, 1.0}. An actual trial with interims (Fig. 14.5) gives rise to a sequence of potential coordinate pairs: $\{(0.2, Z_1), (0.4, Z_2), (0.6, Z_3), (0.8, Z_4), (1.0, Z_5)\}$, (lines connect these points for illustration only) where the Z-values of the primary endpoint are calculated at each of those interim analysis times.

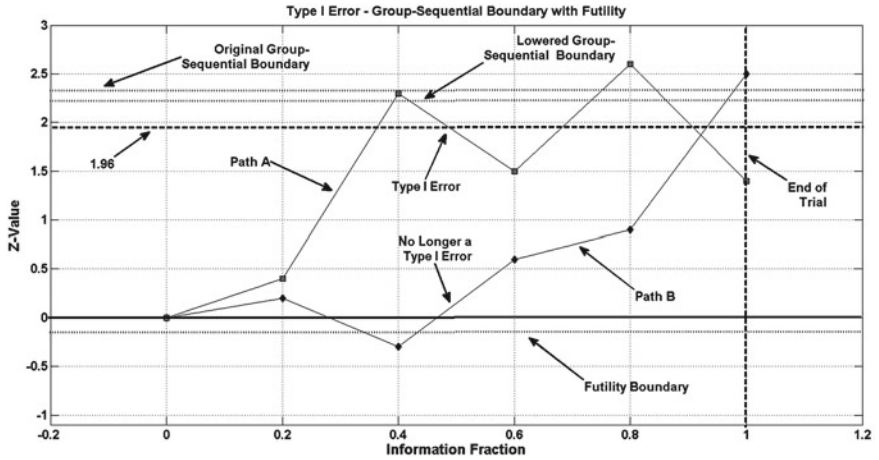


Fig. 14.6 Binding boundaries Example Part 3: group-sequential design with futility boundary

Trial A has squares designating each of the coordinate pairs at the interim analyses; diamonds for Trial B. If significance is claimed at the first interim for which the Z-value exceeds the 1.96 line, it is well-known that the Type I error will be inflated. If 2.41 (the Pocock boundary for 5 looks) is used rather than 1.96, the overall probability of a Type I error for exceeding this boundary is restored to 0.025. Now Trial A crosses the 2.41 line at the 4th interim, and thus is now a Type I error. Trial B remains a Type I error.

The Z-values for Fig. 14.6 are identical to those in Fig. 14.5. Figure 14.6 additionally has a "futility" boundary at $y = -0.2$. For Trial B, the futility boundary is crossed at the 3rd interim analysis, so Trial B should be terminated for futility at information fraction 0.4. Trial B no longer results in a Type I error at the end of the trial because the futility boundary terminates the trial before that Type I error can occur.

So futility boundaries reduce the probability of a Type I error, in this case below 0.025. To compensate, the group-sequential boundary (corresponding, for example, to the horizontal line at 2.41) can be lowered, until the Type I error is restored to 0.025. For example, the horizontal line might drop from 2.41 to 2.36.

Suppose an efficacy boundary has been lowered, as just described, by including a futility boundary in the design. Suppose further, that in the actual implementation of the trial, a crossing of the futility boundary is ignored. Then the lowered group-sequential boundary no longer maintains the Type I error.

When there is a futility boundary, and the group-sequential efficacy boundary has been lowered to compensate, in order to maintain the Type I error, any crossing of the futility boundary must be rigidly adhered to, and the trial stopped if that futility boundary has been crossed. A pair of boundaries consisting of a futility boundary and an efficacy boundary that has had a compensatory lowering are designated a

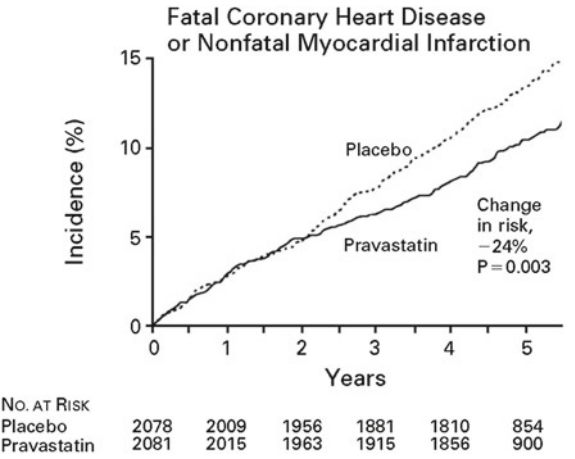


Fig. 14.7 Survival curve from CARE trial—a delayed treatment effect

"binding boundary", because the Type I error is only maintained if a crossing of the futility boundary is binding, requiring termination of the trial.

Such rigidity is difficult if not impossible to enforce. Some statisticians may not fully comprehend the immutable nature of the algorithm. But even if they do, will all those involved in the decision process also understand? For example, consider a statement such as: "agree to terminate the trial if the results are not in the promising zone". Such statements may seem less rigid and eminently reasonable at the kick-off meeting of a DSMB. Assume, optimistically, that the statistician and all other members of a DMSB fully understand the ironclad nature of the rule. DSMBs typically make recommendations that the Sponsor is not bound to follow. Often, a high level executive who is not associated directly with the study may make the final decision. The DSMB generally never meets this person. When a recommendation is made to terminate the trial because the "results are not in the promising zone", but the results do not look too bad to that executive (most futility boundaries can stop a trial with a positive Z-value at some point), will that executive understand the immutable nature of the algorithm and automatically declare futility? The reason people are involved in the decision making, rather than robotically adhering to a calculated number, is because of the tacit understanding that such decisions may involve other considerations.

As an example, the sponsor may decide to carry the trial to full length even if there appears to be no chance that significance will be achieved at the end of the trial. This could happen if the sponsor decides to cease development and sell the drug to a third party for further development. The act of terminating the trial for futility can dramatically reduce the price that could be obtained since the drug will now carry the stigma that "the trial was stopped early for futility". The sponsor may decide to overrule such a futility boundary crossing, and allow the trial to go to natural termination. This is just one of many possible ways in which futility boundaries crossings end up being ignored. This would invalidate the Type I error.

It is important to realize that it is not the actual act of overruling a futility boundary crossing that invalidates the Type I error. It is the *probability* that a futility boundary will be overruled that is crucial for determining the Type I error. And, as the above example reveals, that probability can never be known.

Binding boundaries or variants seemed to have escaped notice for many years or decades in many scholarly and important publications. For group-sequential boundaries, the impact on the Type I error is generally small if the lower boundary is symmetric to the upper boundary in which both are O'Brien-Fleming boundaries with 0.025 1-sided significance level for the lower boundary as well as for the upper boundary. But for those group-sequential boundaries in which the upper and lower meet at the end of the trial, the impact can be substantial. Such boundaries occur frequently in the books of Whitehead (1997), and Jennison and Turnbull (2000). Snapinn (1992), who develops conditional power for proportions in a very useful way, proposed using those conditional power calculations in a binding way ("This paper describes a conditional probability procedure which attempts to maintain the overall significance level by balancing the probabilities of false early rejection and false early acceptance"). In their seminal paper introducing the conditional error function for adaptive designs, Proschan and Hunsberger (1995) take a similar approach stating that the increase in "the Type I error rate can be avoided by … agreeing not to continue the study unless the p value after the first n observations is less than" some prespecified value. They continue by stating that "the idea of reducing the Type I error rate by allowing early termination in favor of [the null] is not new. Gould and Pecore (1982) adopted this approach in a group sequential context." The concept of "Binding Boundaries", in which such approaches were no longer deemed acceptable, appears to have emerged years later. The same basic problem forms the basis for the adaptive methods more recently proposed by Mehta and Pocock (2011)—these are again based on conditional power calculations, this time "agreeing to not continue unless the conditional power is in the promising zone".

14.5 Futility

14.5.1 Futility Boundaries

The preceding section mentioned "futility boundaries", but did not define the concept. Curiously, in perusing the literature of futility boundaries for clinical trials, I could find no mention of a mathematical objective. A Google search led to the following definition of futile: "incapable of producing a useful result". This could be restated statistically as "the probability of a useful result is very low". At the time of an interim of a clinical trial, that probability depends on the data collect thus far, and assumptions regarding the future course during the remainder of the trial. This is actually the definition of conditional power, and the comparison of that power with some quantification of "very low" can be used to assess futility.

But futility boundaries are advocated as an alternative to conditional power. And without a defined goal, or providing any relationship between a given boundary and

the end-trial probability of success, futility boundaries appear to be ad hoc rules. It is sometimes stated that, like efficacy boundaries which preserve the Type I error, futility boundaries preserve the Type II error. Given the null, efficacy boundaries precisely define the critical region and in turn, the probability of a false positive can be calculated precisely. The Type II error is based on an alternative hypothesis which includes an assumed treatment effect. That treatment effect is only a guess, often with little or no data as basis. And it is often a biased guess to accommodate cost and funding as well as other considerations. If the Type II error is wrong at the outset, the rationale for preserving it is not clear. If, at an interim, a futility boundary is crossed, is that a true reflection of a "futile" trial (in the above definition), or of an assumed treatment effect that is far off the mark? With the ability now to increase the trial sample size at the time of an interim, in many cases, a "futile" trial can be transformed into one with good prospects (provided the funding exists to support such an increase). The ad hoc nature of the futility boundary does not provide any quantitative means of making an educated decision as to whether the trial should be terminated for futility; either the futility boundary is crossed, or it is not.

Often, futility boundaries are set up to meet the efficacy boundary at end of trial, the rationale being that if the z-value at the end of the trial is less than, say 2.03 (corresponding to the 5th look under a standard 5-look O'Brien-Fleming efficacy boundary), the trial has failed. But a trial with a final z-value of 1.90 can often be used as supportive, or as mentioned above, as evidence, to a prospective buyer, that a properly designed trial could succeed. The futility boundary is not set up to make exceptions.

The discussion of futility boundaries thus far focused on general properties. In the case of non-proportional hazards, futility boundaries can lead to the wrong decision, and should be avoided. Figure 14.7 from the CARE trial (Sacks et al. 1996) presents estimated KM survival curves for fatal coronary heart disease or nonfatal myocardial infarction. The active treatment is a statin. The overlapping of these survival curves for the first 2 to 2¼ years is indicative of a delayed treatment effect. Similar delays have been observed in survival curves from other statin trials.

Group-sequential and futility boundaries are usually presented in terms of Z-values: Fig. 14.8 gives the O'Brien-Fleming boundary for 5 looks.

Figure 14.9 presents a typical futility boundary and expected z-values from the time-dependent survival curves extracted from the CARE survival curves of Fig. 14.7. Because the recruitment pattern can have a substantial effect on how the expected Z-value evolves with time from study start, expected Z-values are presented for several different recruitment patterns. Expected Z-values assuming an exponential model, rather than the time-dependent failure curves of Fig. 14.7 are also provided for comparison.

Because of the delayed treatment effect, all 3 of the expected Z curves based on the CARE model remain completely horizontal until the end of the delay. By this point, all 3 curves have crossed the futility boundary and remain well below that boundary for most of the trial. Note that the time frame of the expected Z-value is given in calendar time, in which the trial is monitored (i.e., the meetings of the Data Monitoring Committee (DMC) are scheduled in calendar time.) The KM curves of

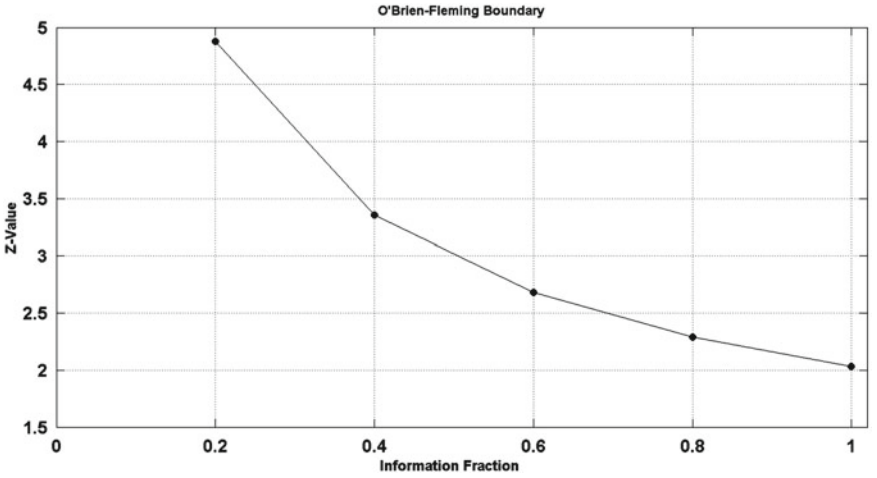


Fig. 14.8 O'Brien-Fleming Boundary for 5 Looks

Fig. 14.7 are given in time from randomization. Had typical futility boundaries been used for monitoring the CARE (or other statin trials) the futility boundaries would undoubtedly have been crossed, with the potential abandonment of this life-saving class of drugs.

Continuing with Fig. 14.9, under the exponential assumption, and uniform recruitment of 3500 patients over 24 months, the expected Z value curve would have remained above the futility boundary until just before the very end of the trial—the futility boundary would have failed to predict the ultimate failure of that trial. In contrast, using the actual time-dependent failure curves from CARE with the same recruitment of 3500 patients over 24 months, the expected Z-value curve crosses the futility boundary with likely early termination for futility. But this curve in the end reaches statistical significance. Suppose, however, with the same initial recruitment pattern of 3500 patients over 24 months, it is decided at the 24-month interim, to increase the sample size by 5000 patients, to be randomized over the period 25–42 months. The trial now fails. The reason is that by adding the massive bolus of patients starting at 25 months, all of the patients starting at or after month 25 will experience the delayed treatment effect until at least 50 months. Thus adding these patients beginning at month 25 will serve to drag down the expected Z-value for most of the remainder of the trial.

Suppose that, instead of enrolling 3500 patients, re-estimating the sample size, and then adding 5000 patients (for which the trial ends in failure), all 8500 patients are enrolled over 24 months. The best result is obtained, with likely early stopping for superior efficacy. But with this scenario, and a futility boundary in place, that futility boundary would be crossed early in the trial, with the expected Z-value remaining below that boundary for most of the trial. Again, early termination for futility is likely.

14.5.2 Conditional Power

Conditional power does not suffer any of the issues discussed above for futility boundaries. However, conditional power calculations are often approached from a too simplistic viewpoint. At first, the following discussion will be restricted to proportional hazards, or more generally, treatment effects that are not time-dependent in non-survival settings.

1. The unreliability of the interim estimate.

At the time of an interim, define conditional power as the probability of end-trial success given the data collected thus far, and given assumptions regarding the remainder of the trial [refer to those assumptions as the Future Treatment Effect (FTE)]. For proportional hazards, the FTE is a single number.

The "data collected thus far" serves two roles:

- In terms of population sampled, it is usually the most representative sample of data from the full data set of the trial that can be obtained (assuming proportional hazards).
- It will be a subset of the final data, and as such, the interim results will have to be "overcome" for the end-trial results to differ meaningfully from those at the interim. The further along the trial, as the ratio of information of pre- to post-interim data increases, the less likely the final analysis will differ meaningfully from that of the interim.

The estimate of efficacy from the data collected thus far will be referred to as the "current trend" [=Current Treatment Effect (CTE)]. Because of the two factors mentioned above, the CTE may appear to be a good choice for basing assumptions for the remainder of the trial.

Still, there is good reason to be skeptical of analyses that blindly use the CTE for the assumed FTE. Take for example, sample size re-estimation. Sample size re-estimation is usually planned because there is concern that an estimate based on the design sample size (see "Design Sample Size" and its confidence interval in Fig. 14.10) lacks the precision to be a convincing estimate. The task is to increase the sample size sufficiently that that re-estimated sample size will provide an estimate with convincing precision. The heuristic rationale for seeking this increased sample size can be envisioned by comparing the widths of the design and re-estimated confidence intervals in Fig. 14.10. At the interim, typically based on 50% or less of the design sample size, the interim estimate has a very large confidence interval. This is an inescapable consequence of the setup depicted in Fig. 14.10. Interim estimates in these situations are highly unreliable.

2. Alternatives to the Current Treatment Effect (CTE).

Figure 14.11 displays a few options for the FTE (the dashed lines beginning at the interim, at month 6). Here, the True Treatment Effect (TTE) is 3. Clearly, the desired choice for the FTE is also 3, the TTE.

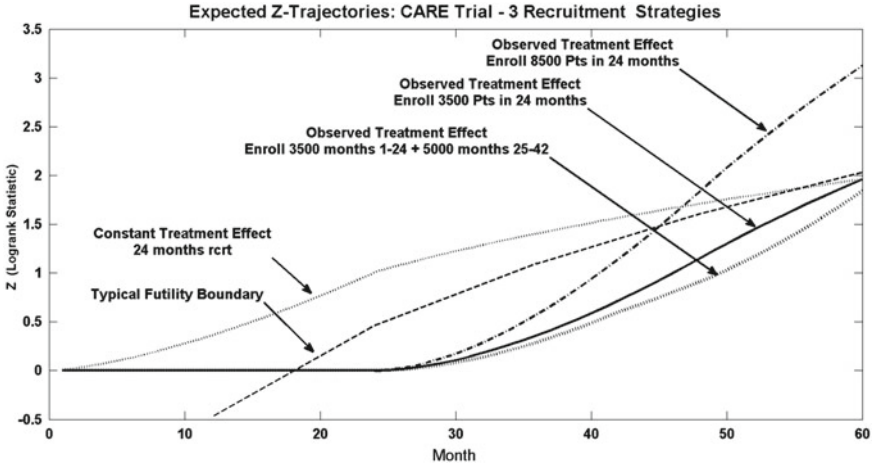


Fig. 14.9 Z-trajectories for CARE trial: the failure of futility boundaries

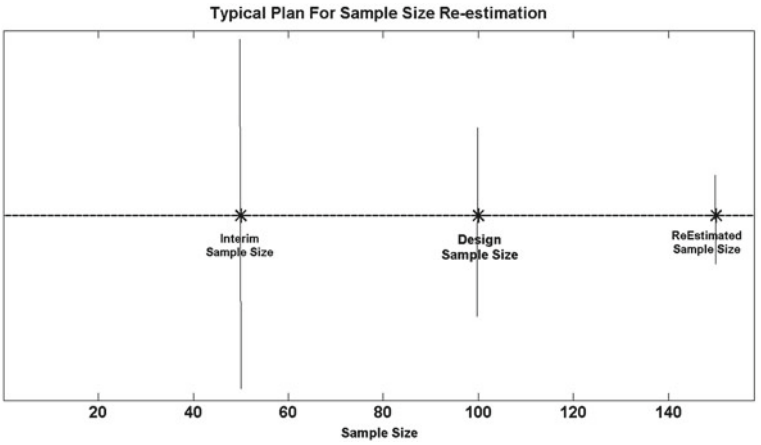


Fig. 14.10 Relative reliability of estimates for sample size re-estimation

Although the CTE is often an unbiased estimate of the TTE, due to its very wide confidence interval, it can easily differ substantially from the TTE. If the CTE (5.5 in Fig. 14.11) is substantially higher than the TTE, then setting the FTE = CTE will compound this error by assuming that the overestimate will continue for the remainder of the trial, an unlikely assumption. This will result in a highly biased estimated conditional power. If the DTE (Design Treatment Effect) is lower than the CTE, there is reason to believe that the CTE is an overestimate.

Rather than relying exclusively on an inherently unreliable estimate, a better strategy is to balance that unreliability against the known weaknesses of the Design Treatment Effect (DTE). Figure 14.11 displays projections based on the CTE alone,

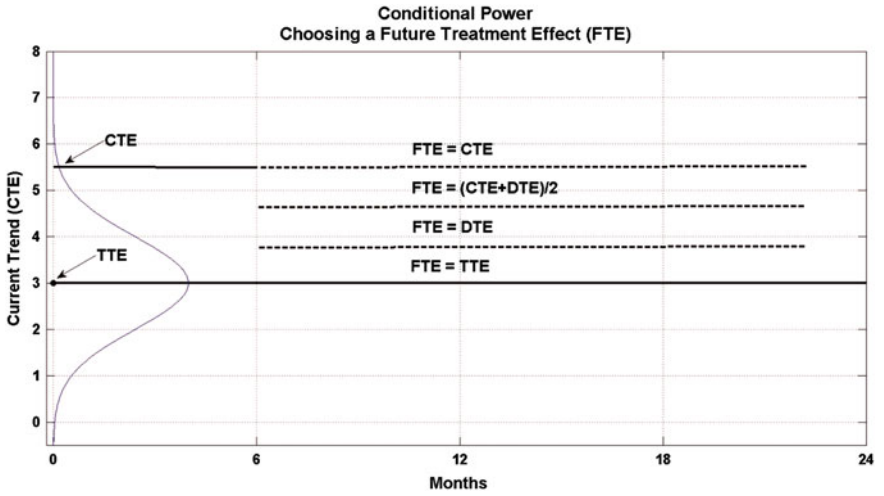


Fig. 14.11 Conditional power: some options for assumptions for the remainder of the trial

the DTE alone, and the average of CTE and DTE. This trio of assumptions underlying conditional power estimates provide a basis for each of the DSMB members (and others who may be involved in the decision making process) to form opinions, based on their individual beliefs regarding the relative strengths of the CTE and DTE. Interim estimates that appear overly optimistic should be tempered. Assumptions other than the dashed lines in Fig. 14.11 are possible.

Arguments similar to the above apply when the CTE is less than the TTE. And similarly, interim estimates that appear overly pessimistic should be tempered.

Mid-trial modifications, such as sample size re-estimation and futility have broad implications. For example, increasing the sample size may convey to investors that the treatment effect is smaller than assumed during the planning stage. This in turn can lead to financial problems. Because of such issues, mid-trial modifications should only be undertaken if the impact is deemed important. In turn, interim trial modifications are likely to occur only if the CTE differs substantially from the DTE, and it is exactly in this situation that the CTE is most likely to be badly biased. Consequently, it is in this situation that one should be most skeptical of assuming that the FTE = CTE.

Some sample size re-estimation procedures rely exclusively on assuming FTE = CTE. This is particularly pernicious with methods that fall within the "binding boundary" concept. The reason is that the Type I error is calculated assuming that FTE = CTE (see, e.g., Mehta and Pocock). Consequently, when there is a binding boundary, the CTE must be used, even if a reasoned judgment would conclude that it is very biased.

14.5.3 Conditional Power When Hazards Are Non-proportional

While the issues discussed above with conditional power when hazards are proportional persist when the hazard are nonproportional, considerably more difficult problems exist for NPH. When hazards are non-proportional, the treatment effect changes with time. Consequently, the horizontal dashed lines in Fig. 14.11 will no longer be straight lines, but rather must reflect the time-dependent nature of the treatment effect. In the very simple case of a delayed treatment effect, the treatment effect has two levels: it is 0 during the no-treatment-effect zone, and $\delta = \text{const} \neq 0$ after that.

In Sect. 14.5.1, the failure of futility boundaries to provide a basis for reasonable decisions when hazards are nonproportional was examined. In this section, difficulties with conditional power in the presence of NHP are discussed. The problems resulting from NPH are very different for conditional power as compared to futility boundaries. With NPH, futility boundaries may fail because there is no known relationship between the crossing of the boundary and the end-trial probability of success. In contrast, conditional power *is* the end-trial probability of success. Problems arise in the calculation of conditional power; one of the most important is the difficulty in obtaining a reasonable interim estimate of the treatment effect for use as the FTE.

Lakatos (2016) provides a detailed explanation of why obtaining an interim estimate of the treatment effect in the simple case of a delayed treatment effect can be so difficult. The CARE trial of Figs. 14.7 and 14.9 provides a good example. In Fig. 14.7, a delay of about 2 years is readily apparent. This, of course, is based on full end-of-trial data. CARE compared a statin with placebo for reducing fatal coronary heart disease or non-fatal myocardial infarction ("MI" will be used to refer to this endpoint) through the reduction of cholesterol. Because this delay has been observed with other treatments that lower cholesterol, the delay appears to be more a function of cholesterol lowering than the specific treatment. It is thought that cholesterol lowering itself does not immediately reduce MI, but rather inhibits plaque accumulation. As the plaque continues to accumulate in the placebo group, eventually there is enough to differentially increase MI in the placebo group. Thus the length of delay has more to do with the placebo group than the active. Certainly, different methods of cholesterol lowering may differentially effect the accumulation of plaque in the active group which will affect the delay as well. Different lengths of delay have been observed.

Suppose, then, that there is interest in designing a trial with a new type of cholesterol lowering treatment. A delayed treatment effect is suspected. But the length of the delay, and even if the treatment is actually effective is unknown—that is why the trial is being performed.

Like the CARE trial, the new trial is planned to last 5 years. An interim for sample size re-estimation after the 3rd year is not of much use. An example of why an increase in sample size after the 3rd year was given in Sect. 14.5.1 (see Fig. 14.9

and the corresponding discussion). With a treatment effect that is delayed 2 years, any patients enrolled after the 3rd year will remain in the no-treatment-effect zone for most of the remainder of the trial, decreasing the power. The remainder of the discussion in this section is limited to interims that take place during the period 24–36 months.

The Cox proportional hazards model will undoubtedly be used to get an estimate of the treatment effect. Even though the proportional hazards assumption is clearly violated when there is a delayed treatment effect, it is used extensively is such settings. The treatment effect that appears on the graph of Fig. 14.7 was obtained using the Cox model.

The expected value of the treatment effect will be 0 (equivalent to a hazard ratio of 1) during the delay. The question is: how long after the delay will it take for the estimated treatment effect to approach a treatment effect that can be expected during the remainder of the trial. The question is complicated by the fact that the treatment effect expected during the remainder of the trial will be determined not only by patients who were randomized early enough that they no longer reside in the no-treatment-effect zone, but also those enrolled more recently, as well as those who have yet to be enrolled. The length of recruitment plays a critical role in the ability to get a reasonable estimate of the treatment effect.

For the Cox model, the ratio of events occurring *after* the no-treatment-effect zone to events occurring *during* the NTE zone is central to that model's ability to provide a reasonable estimate of the treatment effect to serve as the FTE in the conditional power calculation. For an interim performed 36 months after study start, Lakatos (2016) gives an example in which more than 95% of the events occur during the NTE zone, with less than 5% supporting the alternative. In such a situation, the Cox model estimate will be dominated by the 0 treatment effect events. In actuality, it is typically the treatment effect of the alternative that dominates the remainder of the trial, when there is a delayed treatment effect. The main driver of that 95:5 ratio is the recruitment; in that example, the recruitment pattern is nothing out of the ordinary.

The bottom line is that when the delay is around 2 years, and decisions after 3 years are too late, the events generated under the null will dwarf those generated under the alternative, leading to gross underestimation of the treatment effect to be used for the FTE, the treatment effect that is expected to be in effect for the remainder of the trial.

Some useful references for calculating conditional power are Lan and Wittes (1988), and Snapinn (1992).

14.6 Weighted Logrank Statistic

When hazards are non-proportional, the usual logrank statistic is not optimal; in turn, the question of whether to use a weighted version of the logrank arises. A number of weightings for the logrank have been proposed. In this section, we examine a specific

type of non-proportional hazards alternative, the delayed treatment effect, which can provide insight into the broader class of non-proportional hazards.

The CARE trial (Fig. 14.7) exhibits characteristics of a delayed treatment effect: the survival curves coincide for some period, called the no-treatment-effect zone, followed by a separation that appears to be increasing at a constant rate, i.e., proportional hazards is in effect after the no-treatment-effect zone.

Zucker and Lakatos (1990) studied a more general class of alternatives. Rather than restricting the survival curves to coincide during the no-treatment effect zone, there can be gradual separation, with the treatment being fully effective by the end of the no treatment effect zone. After the NTE, the treatment was assumed to be fully effective, i.e., proportional hazards. They also referred to a special class of treatment lags call a "threshold lag", which is the same as a delayed treatment effect.

Focusing on delayed treatment effects offers some advantages, particularly simplicity, over the broader class of general treatment lags. In addition, this restriction may be less important from a practical perspective. It is usually quite easy to identify when survival curves coincide for some period, as in Fig. 14.6, and attribute that coincidence to a period of no treatment effect. In contrast, when survival curves appear to separate slowly, assessing whether the underlying true survival distributions come from either proportional hazards, a delayed treatment effect, or a more gradual separation, may be quite difficult. Distinguishing between these three possibilities in the presence of variability typical in such curves is difficult and error-prone.

Zucker and Lakatos (1990) refer to a weighting that is optimal with respect to a delayed treatment effect ("threshold lag"). That weighting gives 0 weight during the no-treatment-effect zone, and constant weight for the remainder of the trial (during which proportional hazards is assumed). For a delayed treatment effect, define an optimal weighting as one that maximizes the Z-value (ZL use a different definition of optimality). Assume without loss of generality, that a positive Z-value indicates superior efficacy.

It is easy to see why assigning 0 weight to events occurring during the no treatment effect zone, and constant weight during the remainder of the trial maximizes Z. Without loss of generality, assume no ties. Here, as in the calculation of KM estimates of survival, recorded calendar times are converted to time from randomization. Consider the following 2×2 table at the kth ordered event timewhere X_k is

	Event	Event free	Total
Group A	X_k	$m_k - X_k$	m_k
Group B	$1 - X_k$	$n_k - 1 + X_k$	n_k
Total	1	$n_k + m_k - 1$	$m_k + n_k$

the indicator that the event has occurred in group A, m_k and n_k are the numbers at risk just before the kth event in groups A and B, respectively. This table is conditional on the patients at risk just prior to the kth ordered event. The weighted logrank statistic (Mantel 1966; Lakatos 1988; Schoenfeld 1981) is

Table 14.12 Consequences of guessing the length of the delay

Power					
Actual delay	Guessed delay—(diagonal—correct guess)				
Months	0	1	2	2.5	3
	Std LR	Opt-1	Opt-2	Opt-2.5	Opt-3
0	**92.0**	87.1	80.2	77.1	72.9
1	76.7	**85.7**	78.7	75.9	71.2
2	57.2	66.6	**76.9**	73.1	69.0
2.5	47.5	57.0	66.2	**72.3**	67.7
3	38.8	46.7	56.2	61.7	**67.3**

$$Z = \frac{\sum_{k=1}^{d} w_k \left(X_k - \frac{m_k}{m_k + n_k} \right)}{\sqrt{\sum_{k=1}^{d} w_k^2 \left(\frac{m_k n_k}{(m_k + n_k)^2} \right)}} \tag{14.1}$$

Here, $w_k \geq 0$ is the weight corresponding to the kth (conditional) 2×2 table. The term

$$X_k - \frac{m_k}{m_k + n_k} \tag{14.2}$$

in the numerator of (14.1) is in the form of observed minus expected for the kth 2×2 table. For any event occurring in the no-treatment-effect zone, the expected value of the observed minus expected is 0. However, apparent from the denominator, the variance corresponding to this same event is never 0. If a non-zero weight $w_k \neq 0$ is assigned, then the contribution corresponding to the numerator of (14.1) from the kth event will be 0, while the contribution to the denominator will be greater than 0. Consequently, Z in (14.1) will be decreased. The only weighting that will not decrease Z is $w_k = 0$ for all events k occurring in the NTE zone. Proportional hazards is assumed to be in effect for the remainder of the trial, so constant weighting, known to be optimal for proportional hazards is also optimal here. Because w_k^2 appears in the denominator, any constant will do. Gill (1980) provides a more general theory and optimal weighting for the logrank. That weighting, as well as the weighting w_k in (14.1), is a function of the treatment effect, and not the survival function.

The above discussion assumes that the length of the delay is known. But it rarely if ever is known. Unfortunately, the length of the delay is very difficult to estimate, particularly at an interim. One approach is to guess the length of the delay. Table 14.12, based on simulations used in the design of a recent trial, shows the price of guessing incorrectly. In this trial, all patients were to be followed for 1 full year.

In the "0" column, the "guess" is that there is no delay, so the standard logrank is in effect. If the actual delay is also 0, the power is 92%. But if there truly is 1 month delay, the power of the standard logrank drops to 76.7, and if the true is 3 months, the power drops to 39%. So the price, in this study, of ignoring a delayed treatment

effect when there is an actual delay can be quite high. From the other perspective, the price of adjusting for a delay when there is no delay (row "0") is that the 92% power drops to 87% if the weighting is set for 1 month, down to 73% if the weighting is set for 3 months.

The diagonal shows how the power declines even if the guess happens to coincide with the true delay. It may seem surprising that even if one correctly guesses a 3-month delay, the optimal power is 67.3%. The reason is that, as shown above, the optimal weighting gives 0 weight to any event occurring during the no treatment effect zone, so all the events occurring during the first 3 months are lost to the analysis. But, as the argument above shows, if those events are given non-zero weight, the power is further eroded.

Comparing powers off the diagonal to those on the diagonal reveals the price of guessing incorrectly. For example, if the true delay is 2 months, and the guess for weighting is 1 month, then the power drops from an optimal of 76.9 to 66.6. The price of an incorrect guess can be very high.

The above discussion shows the perils of guessing incorrectly. One way to minimize that loss is to optimize for a range. For discussion and derivation of the optimization, see Zucker and Lakatos (1990). In addition to the exact solution, Zucker and Lakatos propose a simplified version, which performs almost as well as the exact solution. That simplified version is very easy to calculate. Consider two statistics: the standard logrank $\left(Z_{\text{std lgrnk}}\right)$ and the logrank statistic calculated on the sorted data beginning at the end of the delay $\left(Z_{\text{post delay}}\right)$. The simplified statistic is

$$Z = \frac{Z_{\text{std lgrnk}} + Z_{\text{post delay}}}{\sqrt{2\left(1+\hat{\rho}\right)}} \tag{14.3}$$

where ρ is the correlation between these two statistics, given by

$$\hat{\rho} = \sqrt{d_{\text{post delay}}/d_{\text{std logrank}}} \tag{14.4}$$

where $d_{\text{std lgrnk}}$ is the number of events for the entire trial, and $d_{\text{post delay}}$ is the number after the delay.

While the statistic in (14.3) is given in the form of a standardized sum of Z-values, it is easy to see that it can be converted into the usual form of a weighted logrank statistic (14.1). The form presented in (14.3) is easier to calculate, since one only has to calculate the Z-value for the usual logrank and the Z-value for the usual logrank restricted to the sorted data whose times exceed t_{delay}, and substitute in (14.3). This calculation can be performed using standard software for calculating the logrank statistic. There is no need for special software to evaluate (14.1).

The critical nature of targeting the weighting to the length of the delay is apparent from Table 14.12. For weightings that are based on the estimated survival curve [e.g., Tarone and Ware (1977), Harrington and Fleming (1982, 1991)], there is no way to target the length of the delay. In fact, the treatment effect, which was shown above

to be the sole determinant of the optimal weighting for a delayed treatment effect, is independent of the survival distribution of the control group.

14.7 Concluding Remarks

This chapter dealt with some standard methods for addressing issues that arise when interim analyses are performed. The presentation critically examines these methods to reveal strengths and weaknesses. The section on Designer Boundaries begins with survival curves from a recently completed trial that was designed using the very standard approach of assuming exponential survival. It develops a group-sequential "Designer" boundary that dramatically reduces the length of the trial and the sample size. It is also "self-adapting" in the sense that there is good power for the trial to end early; if it does not, it can continue to a larger number of events. This Designer Boundary gives rise to a spending function that is not likely to fit within the confines of the many formula-based spending functions. Unfortunately, currently available commercial software for group-sequential methods does not appear to accommodate any variation of spending functions from those formulas.

The next section dealt with "Binding Boundaries", a concept which was popular in the statistical community a dozen years ago, but seems to have disappeared into the nether. The binding boundary concept pointed to flaws in the approach to combinations of efficacy and futility boundaries that inflated the Type I error. Many group-sequential methods proposed prior to the recognition of this concept did indeed incorporate this flaw. But as time has passed, the concept seems largely forgotten, and the flaw again is present in currently-used statistical methods.

The topic of futility was discussed next. The futility boundary appears to be ad hoc, possessing no quantifiable relationship as to whether or not a trial is futile. Conditional power is designed to quantify the probability that a trial will be successful, and should always be chosen in preference to a futility boundary. However, conditional power requires an assumption as to what the treatment effect will be for the remainder of the trial. This is not a weakness of the conditional power concept. It is consequence of attempting quantify the probability that a trial will be successful at some future time. The fact that futility boundaries do not incorporate assumptions about the treatment effect for the remainder of the trial, is not an advantage, but rather the basis for its inability to quantify the end-trial chances of success. The discussion of conditional power challenges the statistician to go beyond simply assuming that the current estimate of the treatment is a good choice for the future treatment effect. For doing so may lead to poor decisions. Alternate approaches were recommended. All of the considerations of this paragraph thus far apply to proportional hazards (i.e., constant treatment effects). When alternatives are nonproportional, the situation is far more complex, and adequate methods for most situations have yet to be developed. In cases when the statistical methodology is questionable, it is usually better to forego making a decision rather than offer one that may be badly flawed.

With nonproportional hazards, the use of a weighted logrank statistic could be considered. One of the simplest forms of nonproportional hazards is the delayed treatment effect. Section 14.6 discusses optimal weighting for a delayed treatment effect. The optimal weighting in this case is also simple, as is the proof of optimality. The weighting depends entirely on the time-dependent nature of the treatment effect, not the form of the survival curve.

References

Bauer, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie, 20*(4), 130–148.

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis*. Hoboken, NJ: Wiley-Interscience.

Gill, R. D. (1980). *Censoring and stochastic integrals*. Amsterdam: Mathematisch Centrum.

Gould, A. L. (1992). Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine, 11*(1), 55–66.

Gould, A. L. & Pecore, V. J. (1982). Group sequential methods for clinical trials allowing early acceptance of Ho and incorporating costs. *Biometrika, 69*(1), 75–80.

Gould, A. L., & Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics-Theory and Methods, 21*(10), 2833–2853.

Halperin, M., Rogot, E., Gurian, J., & Ederer, F. (1968). Sample sizes for medical trials with special reference to long-term therapy. *Journal of chronic diseases, 21*(1), 13–24.

Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika, 69*, 553–566.

Hwang, I. K., Shih, W. J., & DeCani, J. S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine, 9*, 1439–1445. Cited on p. 144.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall.

Kim, K., & Demets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika, 74*, 149–154.

Lakatos, E. (1986). Sample size determination in clinical trials with time-dependent rates of losses and noncompliance. *Controlled Clinical Trials, 7*(3), 189–199.

Lakatos, E., (1988). Sample sizes based on the log-rank statistic in complex clinical trials. *Biometrics*, 229–241.

Lakatos, E. (2002). Designing complex group sequential survival trials. *Statistics in Medicine, 21*(14), 1969–1989.

Lakatos, E. (2015). Optimizing group-sequential designs with focus on adaptability: Implications of nonproportional hazards in clinical trials. In W. R. Young, & D.-G Chen (Eds.), *Clinical trial biostatistics and biopharmaceutical applications* (Chapter 7, pp. 138–178). Boca Raton: Chapman & Hall/CRC.

Lakatos, E. (2016). Sample size for survival trials in cancer. In S. L. George, X. Wang, & H. Pang (Eds.), *Cancer clinical trials: Current and controversial issues in design and analysis* (Chapter 8, pp. 235–277). Boca Raton: CRC Press, Chapman & Hall.

Lan, K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika, 70*(3), 659–663.

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

Lan, K. K., & Zucker, D. M. (1993). Sequential monitoring of clinical trials: The role of information and Brownian motion. *Statistics in Medicine, 12*(8), 753–765.

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

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

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

Perren, T. J., Swart, A. M., Pfisterer, J., Ledermann, J. A., Pujade-Lauraine, E., Kristensen, G., et al. (2011). A phase 3 trial of bevacizumab in ovarian cancer. *New England Journal of Medicine, 365*(26), 2484–2496.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*(2), 191–199.

Proschan, M. A., Follmann, D. A., & Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 1131–1143.

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

Sacks, F. M., Pfeffer, M. A., Moye, L. A., Rouleau, J. L., Rutherford, J. D., Cole, T. G., et al. (1996). The effect of pravastatin on coronary events after myocardial infarction in patients with average cholesterol levels. *New England Journal of Medicine, 335*(14), 1001–1009.

Shih, W. J. (1993). Sample size reestimation for triple blind clinical trials. *Drug Information Journal, 27*(3), 761–764.

Snapinn, S. M. (1992). Monitoring clinical trials with a conditional probability stopping rule. *Statistics in Medicine, 11*(5), 659–672.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika, 68*, 316–319.

Self, S. G. (1991). An adaptive weighted log-rank test with application to cancer prevention and screening trials. *Biometrics*, 975–986.

Tarone, R. E., & Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika, 64*, 156–160.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. Chichester: Wiley.

Wittes, J. & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine, 9*(1–2), 65–72.

Yang, S., & Prentice, R. (2010). Improved logrank-type tests for survival data using adaptive weights. *Biometrics, 66*(1), 30–38.

Zucker, D. M., & Lakatos, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika, 77*, 853–864.

Chapter 15
On Design and Analysis of Dose-Response Trials for Early Clinical Development

Qing Liu

15.1 Introduction

15.1.1 Regulatory Principles

To develop a new drug for non-life threatening medical conditions, conducting placebo controlled dose response trials at the early stage of a clinical development may reduce the number of failed phase 3 trials. It is essential that an early trial provides

I. dose-response information for short term desirable and potentially undesirable effects, and
II. a reliable estimate of the full dose range to carry forward to phase 3 dose-response trials.

Objective I plays a critical role in decisions to continue the clinical development. This objective is well served for trials which employ the triple trends test developed by Capizzi et al. (1992) for efficacy analysis and early work by Tukey et al. (1985) for safety analysis.

For new drugs with potential long-term safety concerns, objective II is vital for the success of phase 3 trials and post-approval patient care. It is all too common to discover that an approved drug has a high long-term safety risk, that regulatory agencies have to withdraw marketing approvals or to require black-box safety warning restrictions. Occasionally, regulatory agencies would approve a drug with questionable efficacy and issue black-box safety warning restrictions to limit the drug's usage only to patients who may have failed other drugs. Because early dose response trials are often designed with short or intermediate treatment and follow-up, the results of

Q. Liu (✉)
Quantitative & Regulatory Medical Science, LLC, Long Valley, NJ 07853, USA
e-mail: Liu.Qing@QRMedSci.net

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3_15

early trials cannot reliably predict the phase 3 dose with optimal long term benefit-to-risk profile. Rather, the early trials should be designed and analyzed in such a way that identifies a reliable *lower effective bound* (LEB) for effective doses, which along with a lower less than fully effective dose and the highest short-term safe dose form the full dose range for carrying forward to phase 3 clinical development programs.

While the importance of phase 3 dose-response trials has not been broadly incorporated in clinical development programs by the pharmaceutical industry, the consequence of limited phase 3 dose-response information has been well recognized by regulatory agencies. Temple (2004) states that "*The impression that dose-finding is largely completed in phase 2 is a terrible error. Phase 2 studies almost never can detect small differences in effect, and cannot give useful information on safety except for the most common events.*" Temple (2004) further suggests to "*study a full range of doses in phase 3 to establish dose response for both favorable and unfavorable effects and to locate less than fully effective dose that may still be useful.*" The need for phase 3 dose response trials is also clearly elaborated by Hemmings (2006, pp. 30, 46–47), who details various clinical issues and regulatory ramifications of selecting one dose for phase 3 trials. In particular, Hemmings (2006, p. 47) states that "*Where data on outcome are required for submission, it is considered that the continuation of dose-finding into phase III would usually be highly beneficial, using phase II trials with a surrogate only to narrow the potential dose range rather than to select a single dose for the phase III study.*"

15.1.2 Motivation

For the design and analysis of early dose-response trials, these regulatory principles have been traditionally addressed by step-down trend tests. In particular, Quan and Capizzi (1999), following Tukey et al. (1985), apply a triple trends test in a step-down fashion to identify a no-statistical-significance-of-trend (NOSTASOT) dose for a two-way dose-response designs. They show that this step-down trend test is more powerful than a commonly used step-down ANOVA test. The triple trends test is also robust against various shapes of the dose-response relationship. For objective II, the step-down trend test has been successfully applied for phase 2 as well as phase 3 dose-response trials.

It is known that a trend test is limited only to doses that are studied in a trial. In addition, the step-down trend test reduces to a pair-wise test for the lowest dose studied in the trial. Thus, the power for testing the lowest dose is substantially reduced. For many applications, inference of treatment effect for doses not studied in the trial is also important. For example, suvorexant was shown effective for patients with insomnia with 20/15 and 40/30 mg regimens in phase 3 trials. Because of safety concerns of the 40/30 mg regimen, the U.S. Food and Drug Administration (FDA) advised that the efficacy of 10 mg is also established, that 10 mg should be the starting

dose for most patients, and furthermore, a 5 mg dose would be necessary for patients taking concomitant moderate CYP3A4 inhibitors. The 10 mg dose was not studied in phase 3 and was considered ineffective by the sponsor from a phase 2 cross-over study, while the 5 mg dose does not appear to have been studied at all in the clinical program. One potential approach for inference on a dose that has not been studied is through the traditional pharmacokinetics (PK) and pharmacodynamics (PD) dose-response modeling approach proposed by Sheiner et al. (1989). With dose-response modeling, such inference is conceptually feasible through interpolation within the dosing range of the study.

Note that this *interpolative* inference is different from another use of the dose-response modeling to identify a *minimum effective dose* (MED) for an effect size, say δ, of interest. For example, Lockwood et al. (2003) study the performance of a dose-response modeling approach for estimating the MED. They report that "*the identification of the selected dose-response feature with any real precision from the trial design paradigm is borderline.*" Furthermore, "*The marginal precision raises the question as to what is the best dose to study to ensure a clinical outcome of at least a one-point change in pain score, given the dosing options available.*" At the fundamental level, the statistical idea of MED, which are commonly attributed to Ruberg (1989), does not work well with dose-response modeling in clinical pharmacology. When the maximum treatment effect $\Delta_{\max}$ attributable to the drug is below the specified effect δ, then the MED does not exist. Thus, the MED for a given non-zero δ is not an *intrinsic characteristic* of the dose response model. In contrast, an efficacy threshold (e.g., ED_{90}, the lowest dose at which the dose-response curve achieves 90% of the maximum efficacy $\Delta_{\max}$) is an intrinsic characteristic of interest.

A potential issue of relying only on the models of a triple trends test is the ability to correctly estimate the dose-response relationship. This happens when there is a discrepancy between the shape dose-response curve and the three trend scales of Tukey et al. (1985). The problem is addressed in the two-stage adaptive dose-response design by Liu and Pledger (2005) where a sigmoidal $E_{\max}$ model is used to fit the first stage data. The fitted model is used to derive an adaptive trend test statistic for the second stage and then combined with the first period pair-wise test statistic. The combination trend test is applied in a step-down fashion to identify a NOSTASOT dose. By construction, their combination test is robust and has the ability to identify an effect size of interest with high probabilities over a broad range of models (see Tables 1 and 2 of Liu and Pledger 2005).

A third and more fundamental issue is that a step-down test procedure inherits the classical difficulty of hypothesis testing, which is the lack of concordance between statistically significant and clinically meaningful difference. A statistically significant difference may not be clinically meaningful; a clinical meaningful difference may not be statistically significant.

To our best knowledge, there is no procedure that resolves these problems for parallel group dose-response designs. These are the motivations of the research.

15.1.3 Approach

We consider the classical parallel group design where patients are randomized to receive a placebo or one of several doses of an investigational drug. We consider in Sect. 15.2 a general framework where the clinical endpoint in question follows a distribution from a broad families of distributions. This permits practical applications in clinical trials where efficacy (primary and secondary) and safety endpoints have different types of distributions. We assume that the canonical link follows a nonlinear dose-response model (e.g., sigmoidal $E_{\max}$ model), and apply the three-step algorithm for obtaining maximum likelihood estimates (MLEs) of the model parameters considered for the adaptive procedure in Liu and Pledger (2005). This algorithm takes advantages of existing software packages for a wide variety of linear models, e.g., ANOVA, generalized linear models, Cox's proportional hazards model, linear mixed models, and generalized linear mixed effect models (Breslow and Clayton 1993; Liu and Pierce 1993). The algorithm is also applicable to settings where the distribution of the endpoint is unknown but regression analysis can be performed via generalized estimating equations (GEE) models of Liang and Zeger (1986). For objective I, we develop a generalized multiple trends test to establish dose-response relationship. We also develop a unified approach for sample size calculation based on multiple trends test to ensure robust power where the canonical link is described by a general dose-response model. For objective II, we propose a likelihood inference of the effect of any given dose, which may not be used in the actual trial. The three-step algorithm is used to obtain point estimates of effect size. We then apply a bootstrap method to the algorithm to obtain the confidence intervals. We construct a likelihood test with superiority margins that are calibrated to achieve the specified minimum power for detecting an effect size of interest. We refer this as the likelihood *test for statistical significance and clinical meaningfulness* (TSSCM), and apply it to identify the *lower effective bound* (LEB) with the specified probability of coverage of the dose with the effect size of interest. The LEB is then used to define the *full dose range* for phase 3 trials according to the development strategy by Temple (2004). A key feature of the strategy is to study "*whether sub-effective dose represents some people responding fully or all people responding a little*" (Temple 2004). We describe in Sect. 15.5 simulation study results to confirm that the proposed methods meet the specified probability criterion. In addition, we evaluate the probability of success of the development strategy of Temple (2004) with the proposed LEB approach.

15.1.4 Research History

The methods described in this paper were developed by the author in the context of the PhRMA working group on dose-finding. The work was presented at the Joint Statistical Meeting (JSM) in 2006. Both this paper and the 2006 JSM presentation use the identical core three-step algorithm for modeling fitting, which was also considered

for the adaptive procedure in Liu and Pledger (2005). With the introduction of the superiority margins, the TSSCM is used in this paper to define LEB.

Another difference is the example used for numerical illustrations. The case example in osteoarthritis of this paper was not mature in 2006, and thus, an example in neuropathic pain was used for the 2006 JSM presentation. The key goal of the author's PhRMA working group research was to develop methods with the ability to identify an efficacy threshold and a full dose-range for carrying over to phase 3 clinical development. This was in contrast to the "target-dose" estimation scheme, as described in the PhRMA working group reports (Bornkamp et al. 2007; Dragalin et al. 2010; Pinheiro et al. 2010). The problem with the target-dose scheme was pointed out to the working group following a literature review, which cites work by Lockwood et al. (2003) who conclude that "*the identification of the selected dose-response feature with any real precision from the trial design paradigm is borderline. Therefore, if the objective was to confirm the outcome in a future phase 3 study, selecting a dose based on this single outcome might be 'risky'*." Thus, the author's research goal was apparent given in addition the knowledge of regulatory principles for drug development gained from his prior employment at the FDA, and later industry working experience in trial design and analysis, and licensing and acquisitions of new drugs with optimal benefit-risk profiles to address unmet medical needs.

The author was extremely familiar with many clinical development programs of cyclooxygenase-2 (COX-2) inhibitors, and in particular, was interested in the case example with lumiracoxib for osteoarthritis. At the time the PhRMA working group was formed in 2005, the drug showed at high dose an increased rate for serious liver abnormalities with the hazard ratio 1.92 (Schnitzer et al. 2004). This author hypothesized during that time that the optimal development plan would be to study the two lowest doses (50 and 100 mg bid) in phase 3 program, rather than the higher doses (200 mg bid and 400 mg od). The drug was later withdrew from major markets in 2007 following eight serious liver adverse events, including two deaths and two liver transplants (Hinz et al. 2009). After the withdrawal, additional long term safety data of the 100 mg od or bid dose (Sheldon et al. 2008; Fleischmann et al. 2008) became available. In the dose-response trial in volunteers (Hinz et al. 2009), it was shown that a single dose 50 mg lumiracoxib has a comparable blood monocyte COX-2 inhibitory profile to a single dose of 100 or 200 mg. This suggests that doses lower than 100 mg may be sufficient for pain management. By the CHMP comprehensive assessment report in 2011, it is clear that the author's initial hypothesis of the optimal development plan was well founded, and following this plan the uncertainty concerning the residual risk of hepatotoxicity of the 100 mg od dose may be better addressed. Given the maturity of the totality of data from this case example, we now present the proposed LEB approach and its illustration.

Recently, a colleague brought to my attention of the fascinating article by Parker (2013) of the New Yoker on suvorexant for insomnia. As mentioned early, the suvorexant example is used to illustrate the need for interpolative inference. More importantly, the example reinforces the importance of the regulatory principle (Temple 2004) and show that the proposed LEB approach can in fact fulfill the promise of properly guiding the phase 3 trial designs. Another addition is the bevacizumab

example for proposal of a biphasic dose-response model for which hormesis may be justified on biological ground.

15.1.5 Outline

We provide design and analysis of phase 2 dose-response trials. The main methods are developed in Sects. 15.2, 15.3 and 15.4. We illustrate the proposed LEB approach in Sect. 15.5. As a comparison, we also examine the MED or "target-dose" scheme proposed in the PhRMA working group reports (Bornkamp et al. 2007; Dragalin et al. 2010; Pinheiro et al. 2010). In Sect. 15.6 we discuss several topics, including the lack of intrinsic quality of the MED, the law of parsimony in the sigmoidal $E_{\max}$ model, and the determination of phase 3 full dose-range for both the lumiracoxib and suvorexant case examples. We also suggest areas of future research.

15.2 Dose-Response Models

15.2.1 Basic Design

Let patients be randomized to receive one of the $K+1$ increasing doses $d_0 < d_1 < \cdots < d_K$, where $d_0 = 0$ is the dose for the placebo group. Let n_k be the number of patients randomized to receive dose d_k for $k = 0, 1, \ldots, K$. For a given endpoint, let δ_k be the treatment difference between the effects of dose d_k and placebo (i.e., $d_0 = 0$). It is important to choose δ_k on the canonical scale of the underlying distribution so that it can be easily estimated via existing regression methods widely implemented in commercially available software packages. For the dose response study, it is assumed that δ_k follows the nonlinear model

$$\delta_k = \Delta_{\max} f(d_k; \underline{\nu}), \tag{15.1}$$

where $\Delta_{\max}$ represents the maximum effect size of treatment with the drug, and $f(d_k; \underline{\nu})$, for $k = 1, 2, \ldots, K$, is a nonlinear shape function with range $[0, 1]$ indexed by a vector of shape parameters $\underline{\nu}$.

In this paper, we will present an design to achieve the following seemingly contradictory goals that inference of dose-response must be robust against uncertainties in dose-response relationship and the sample size be sufficiently small.

15.2.2 Characteristics of Dose-Response Curve

It is easy to systemically describe a dose-response curve as a continuous function

$$\Delta(d) = \Delta_{\max} f(d; \underline{\nu}), \tag{15.2}$$

of the dose on a continuous range $d \geq d_0 = 0$. For a given dose d, which may not be formulated for clinical trials or marketing approval, $\Delta(d)$ is the difference of treatment effect on the canonical scale between dose d and dose $d_0 = 0$. By definition, $\Delta(d_0) = 0$.

A monotone shape function $f(d; \underline{\nu})$ is adequate for most small molecular drugs. From clinical pharmacology, a standard dose response curve, in terms of the treatment difference model in Eq. (15.2), is defined by three parameters: the maximum effect $\Delta_{\max}$, the dose, $d_{E_{50}}$, at which $\Delta(d_{E_{50}}) = \Delta_{\max}/2$, and the slope ρ (i.e., tangent) of the curve at $d_{E_{50}}$. The maximum effect $\Delta_{\max}$ is used to describe the maximum *efficacy* attributable to the drug when the dose goes to infinity. Dose $d_{E_{50}}$ represents the *potency* of the drug; a drug with a smaller $d_{E_{50}}$ has a higher potency than a drug with a larger $d_{E_{50}}$.

A parsimonious model that provides the systemic description of dose-response relationships is the sigmoidal model

$$f(d; \underline{\nu}) = d^{\rho}/(d^{\rho} + d_{E_{50}}^{\rho}), \tag{15.3}$$

for $\underline{\nu} = (d_{E_{50}}, \rho)$, which is a pharmacodynamic model derived from a receptor theory (Wagner 1968). Because of its biological basis and empirical appeal of parsimony, the sigmoidal model has been widely accepted in clinical pharmacology (Lalonde 1992). The parsimonious nature also leads to simplification of statistical analysis. For example, Thomas (2006) demonstrates that "*the basis functions* (Bretz et al. 2005) *can be closely matched by the expanded* $E_{\max}$ *model, so the use of the single expanded model does not practically restrict their choice of contrasts.*" See Fig. 2 of Thomas (2006). Thus, for monotone dose-response problems, the small subset of *ad-hoc* models by Bretz et al. (2005) are previously used by Sheiner et al. (1989). Without the distraction of these seemingly unrelated models and the needless multiple contrast test, research could now focus on the fundamental issues of inference relating to the sigmoidal $E_{\max}$ model.

Both the slope ρ and dose $d_{E_{50}}$ define the *threshold*, for either efficacy or safety, at a given percentage level (e.g., 90% or 5%). The determination of efficacy and safety thresholds are important for characterizing benefit-risk profiles of a dose. However, it is well known that estimation of $d_{E_{50}}$ can be extremely difficult (Sheiner et al. 1989, 1991; Kirby et al. 2011). For the suvorexant example, analysis using an $E_{\max}$ model yields a statistically non-significant results for latency to onset of persistent sleep (LPS), which contradicts the fact that the 10 mg dose is already at the plateau of the dose-response in LPS.

Liao and Liu (2009) propose a 5-parameter model used in bioassays that includes the sigmoidal $E_{\max}$ model as a special case. Their model permits asymmetry around $d_{E_{50}}$ to better reflect the underlying biological processes. This model can be applied to model dose-response when deviations from the sigmoidal $E_{\max}$ model are expected on biological ground.

15.2.3 Biphasic Dose-Response

There are, however, increasing number of cases to support the presence of biphasic dose-response relationships (Reynolds 2010). This is especially the case for biologic drugs. For example, in an early phase 2 study Kabbinavar et al. (2003) report that bevacizumab shows a high tumor response rate at 5 mg/kg than 10 mg/kg. The explanation is that the reduced effect at high dose may be due to suppressed growth of new vessels carrying the drug to the tumor (Javaherian et al. 2011). To assist trial design or analysis, we propose an empirical shape function

$$f(d;\underline{\nu}) = d^{\rho_1}/(d^{\rho_1} + D_1^{\rho_1}) - \tau d^{\rho_2}/(d^{\rho_2} + D_2^{\rho_2}). \tag{15.4}$$

The five parameter shape vector $\underline{\mu} = (\rho_1, D_1, \tau, \rho_2, D_2)$ provides a family of mixtures of sigmoidal and reverse sigmoidal curves. As the model includes the sigmoidal curve as a special case (i.e., setting $\tau = 0$), it may have some practical advantages over the non-monotone model proposed by Sheiner et al. (1991).

15.2.4 Discretized Maximum Likelihood Estimate

The nonlinear model in Eqs. (15.3) and (15.4) fall into the general nonlinear models by Davidian and Giltinan (1995, pp. 55–56), for which the parameters of the model could be estimated by a generalized least squares (GLS) method. In our experience, however, it is difficult to obtain consistently the maximum likelihood estimates of all parameters. It is especially problematic for many pharmacologically motivated models (e.g., sigmoidal $E_{\max}$ model) for which the likelihood functions are ill-behaved for the nuisance vector $\underline{\nu}$ with small $\Delta_{\max}$ such that the underlying Newton-Raphson algorithm of the GLS method often fail to converge. In large simulation studies the GLS method could fail to converge for a fraction of simulation runs (Kirby et al. 2011), and thus, wasting the time and resource committed to the simulation studies. The problem also occurs often in safety analysis of low event rates with sigmoidal $E_{\max}$ model fitted on the logit (i.e., log odds-ratio) scale.

To resolve these difficulties, a three-step algorithm for fitting the first stage dose-response data of two-stage adaptive design to a sigmoidal $E_{\max}$ model was developed for use in Liu and Pledger (2005). The goals of the three steps are to obtain the unconstrained MLE of δ_k for $k = 1, 2, \ldots, K$, the least squares estimate of $\Delta_{\max}$, and the best-fitting nuisance vector $\underline{\nu}$. We now present the details of the the algorithm.

Unconstrained MLE of δ_k

The first step is straight forward and does not involve development of new statistical methodologies. The underlying patient level data can be analyzed via an appropriate linear model, e.g., ANCOVA or logistical regression analysis, that includes treatment contrasts between dose d_k and the placebo (i.e., $d_0 = 0$), for which the coefficients of

the contrasts are δ_k for $k = 1, 2, \ldots, K$. Where it is appropriate, the model may also includes baseline prognostic factors as well as study centers. From such an analysis, we obtain the MLE $\hat{\delta}_k$ for $k = 1, 2, \ldots, K$ and their variance-covariance matrix $\hat{\Sigma}$.

There are simple situations (e.g., lumiracoxib trial by Schnitzer et al. (2004) that the first step can be based on summary statistics extracted from published literature. For normally distributed data, when the summary statistics of mean and standard deviation are available, one can calculate $\hat{\delta}_k$ for $k = 1, 2, \ldots, K$. Following the design of the trial, it is possible to construct the variance-covariance matrix $\hat{\Sigma}$. Thus, it is possible to evaluate alternative design and analysis without the source data.

Least Squares Estimate of $\Delta_{\max}$

For a given nuisance vector $\underline{\nu}$, let $S_k = f(d_k; \underline{\nu})$ for $k = 1, 2, \ldots, K$. We call $\underline{S} = (S_1, \ldots, S_K)^T$ the shape vector as it describes the shape of the dose response relationship. Let $\underline{\hat{\Delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_K)^T$ be written as following the equation

$$\underline{\hat{\Delta}} = \underline{S}\,\Delta_{\max} + \underline{e},$$

where the residual vector $\underline{e}$ is assumed to follow an asymptotic normal distribution with a zero mean vector and variance-covariance matrix $\hat{\Sigma}$. The the least square estimate of $\Delta_{\max}$ is given by

$$\hat{\Delta}_{\max} = \underline{S}^T \hat{\Sigma}^{-1} \underline{\hat{\Delta}} / \underline{S}^T \hat{\Sigma}^{-1} \underline{S}, \tag{15.5}$$

for which the estimate of the variance is

$$\text{Var}\,(\hat{\Delta}_{\max}) = 1/\underline{S}^T \hat{\Sigma}^{-1} \underline{S}. \tag{15.6}$$

Best-fit Estimate of $\underline{\nu}$

Instead of using a continuous space, let the vector of nuisance parameters $\underline{\nu}$ be restricted to a carefully chosen fine grid of discrete values. For any given $\underline{\nu}$ from the grid, we consider test against the null hypothesis $H_0 : \Delta_{\max} \leq 0$ in favor of the alternative hypothesis $H_A : \Delta_{\max} > 0$ by the test statistic

$$Z = \{\text{Var}\,(\hat{\Delta}_{\max})\}^{-1/2} \hat{\Delta}_{\max} = \{(\underline{S}^T \hat{\Sigma}^{-1} \underline{S})^{-1/2} \underline{S}^T \hat{\Sigma}^{-1}\} \underline{\hat{\Delta}}. \tag{15.7}$$

The test statistic Z implicitly depends on the nuisance vector $\underline{\nu}$. We then maximize Z over the grid. This results in the discretized maximum likelihood estimate $\underline{\hat{\nu}}$ of the nuisance vector $\underline{\nu}$. Such estimate provides a "best fit" of the data by the corresponding model $\hat{\Delta}_{\max} f(d_k; \underline{\hat{\nu}})$ over all possible models defined by $\underline{\nu}$ over the grid. For sigmoidal curves given in Eq. (15.3), $\underline{\nu} = (d_{E_{50}}, \rho)$, and thus, the best-fit model could be identified over a carefully constructed two-dimensional grid of $d_{E_{50}}$ and ρ.

Note that the shape vector $\underline{S} = (S_1, \ldots, S_K)^T$ can also be based on other nonlinear models such as the biphasic dose-response model given in Eq. (15.4).

15.3 Multiple Trends Test

15.3.1 Maximum Effect

For the endpoint in question, objective I can be formulated as a testing problem against the null hypothesis $H_0: \Delta_{\max} \leq 0$ in favor of the alternative hypothesis $H_A: \Delta_{\max} > 0$ at the significance level α (Sheiner et al. 1989).

For monotone dose-response relationship, we use the triple trends test with trend scales based on sigmoidal curves by Eq. (15.3). For biphasic dose-response relationship, we may select up to five curves according to Eq. (15.4). In this paper, we only address design and inference with methods that use monotone dose-response models. However, through simulation studies, we evaluate the robustness of the design and inference when the response data are generated with biphasic models. This is important as in practice with clinical evaluations, it is possible to observe dose-response relationship with reversal effects at high dose even though there is no biological reason to support a biphasic dose-response relationship.

15.3.2 Choice of Triple Trend Scales

For monotone dose-response relationship, scales for the triple trends test are based on three sigmoidal curves. To directly reflect the notion of the *threshold* effect, we consider an alternative parameterization of the shape function in Eq. (15.3):

$$f(d; \underline{\nu}) = \frac{d^{\rho}}{d^{\rho} + d_{E_\eta}^{\rho}(1/\eta - 1)}, \tag{15.8}$$

where η is the threshold parameter (e.g., 90% or 10%) and d_{E_η} is the dose at which $\Delta(d_{E_\eta}) = \eta\, \Delta_{\max}$. For many applications, the dose at which the efficacy starts to plateau, which is commonly defined as a 90% of the maximum efficacy $\Delta_{\max}$ with $\eta = 0.9$, is of considerable interest. With a given η, the shape vector is $\underline{\nu} = (d_{E_\eta}, \rho)$.

We choose three shape vectors to define the trend scales of the triple trends test. This results in three dose-response curves which are referred to as top, middle and bottom curves. The top curve is selected to ensure detecting the dose response when that the lowest dose d_1 hits the threshold η, i.e., $d_{E_\eta} = d_1$. The bottom curve is selected to reflect the other extreme that the highest dose d_K hits the threshold for which $d_{E_\eta} = d_K$. The middle curve is selected based on clinical input where a dose d_k between d_1 and d_K is set to d_{E_η}. The slope parameter is set to be ρ_T and ρ_B for the top and bottom curves, respectively. As will discussed further below, the slope parameter for the middle curve is calculated to ensure a minimum power criterion for the triple trends test is met. Figure 15.1 plots three dose response curves with the threshold effect size $\Delta(d_{E_{90}}) = 11$ for the lumiracoxib case example in Sect. 15.5.

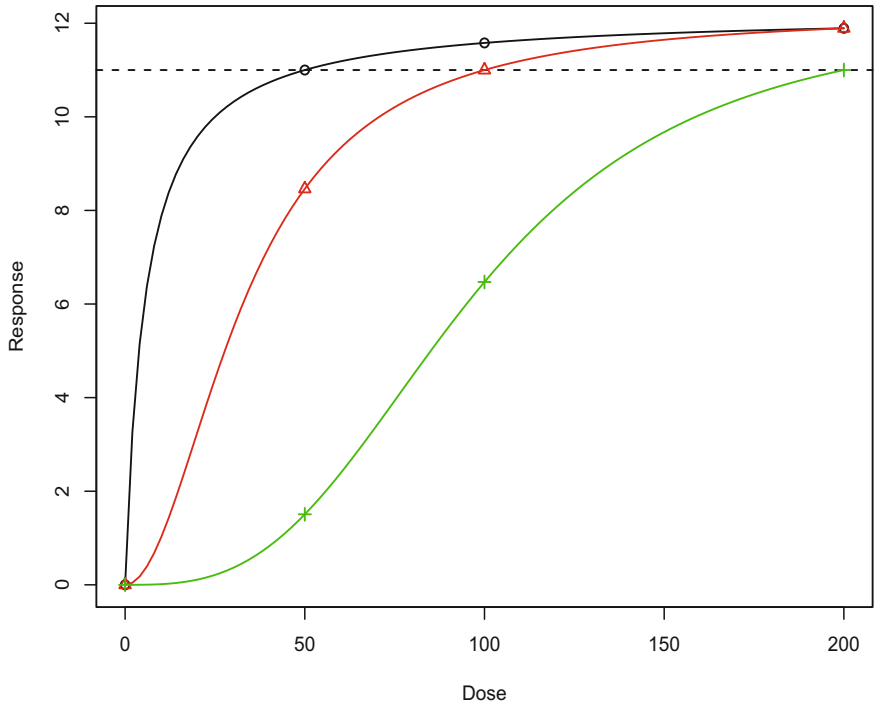


Fig. 15.1 Sigmoid $E_{\max}$ dose response curves

15.3.3 Triple Trends Test

For the ith curve for $i = 1, 2, 3$, the trend test statistic following Eq. (15.7) is

$$Z_i = \{(\underline{S_i}^T \hat{\Sigma}^{-1} \underline{S_i})^{-1/2} \underline{S_i}^T \hat{\Sigma}^{-1}\} \hat{\underline{\Delta}}. \tag{15.9}$$

where the underlying trend scales are based on the shape vector $\underline{S_i}$. A robust test against H_0 in favor of H_A is based on the triple trends test with

$$Z^* = \max\{Z_1, Z_2, Z_3\} \tag{15.10}$$

For a significance level α, the critical value z_α^* is chosen to satisfy

$$P_{M_0}\{Z^* \geq z_\alpha^*\} = \alpha, \tag{15.11}$$

for which $P_{M_0}\{\cdot\}$ is the probability of an event under a null model M_0 with $\Delta_{\max} = 0$. Under the null model M_0, the trend test statistic Z_i follows an asymptotic standard normal distribution. For two different trend statistics Z_{i_1} and Z_{i_2}, their correlation is approximately

$$\text{Cov}\,(Z_{i_1}, Z_{i_2}) = (\underline{S}_{i_1}^T \hat{\Sigma}^{-1} \underline{S}_{i_1})^{-1/2} (\underline{S}_{i_1}^T \hat{\Sigma}^{-1} \underline{S}_{i_2}) (\underline{S}_{i_2}^T \hat{\Sigma}^{-1} \underline{S}_{i_2})^{-1/2}.$$

The type 1 error rate of the triple trends test $Z^* \geq c$ for any critical value c can then by evaluated (Capizzi et al. 1992; Genz 1992). Let z^* be the observed triple trends test statistic, then the p-value of the triple trends test is

$$p^* = P_{M_0}\{Z^* \geq z^*\}. \tag{15.12}$$

We reject the null hypothesis H_0 in favor of the alternative H_A if $z^* \geq z_\alpha^*$ or equivalently if $p^* \leq \alpha$.

15.3.4 Sample Size Calculation

To assess the power of the triple trend test, it is necessary to specify the maximum effect $\Delta_{\max}$ as well as the shape vector $\underline{\nu} = (d_{E_\eta}, \rho)$. This is achieved as follows. For an effect size δ of interest, we assume that there is a dose $d_\delta \geq d_1$ such that $\Delta(d_\delta) = \delta$. We then require that the dose-response curve at d_δ reaches the threshold η, i.e., $d_\delta = d_{E_\eta}$. This leads to the equation

$$\Delta(d_{E_\eta}) = \Delta_{\max}\eta = \delta,$$

from which we have the required maximum effect is $\Delta_{\max} = \delta/\eta$. To calculate the required sample size to achieve a given power $1 - \beta$, say, $\beta = 0.05$, we consider the worst case scenario with the bottom curve of the sigmoidal model with $d_{E_\eta} = d_K$ and its shape parameter ρ_B. Let M_δ be the fully parameterized dose-response model.

The expectation of Z_i under M_δ for $i = 1, 2, 3$ can then be derived. The asymptotic variance-covariance matrix of (Z_1, Z_2, Z_3) remains unchanged. Thus, the power of the triple trends test $P_{M_\delta}\{Z^* \geq z_\alpha^*\}$ can be evaluated.

As mentioned early, the slope ρ of the sigmoidal model for the middle curve is yet to be determined, resulting in a parameterized trend statistic $Z_2(\rho)$ for the middle curve. Let $Z^*(\rho) = \max\{Z_1, Z_2(\rho), Z_3\}$, we require that the sample size n_k for $k = 1, 2, \ldots, K$ to satisfy the following equation

$$\min_{\rho \in [\rho_1, \rho_2]} P_{M_\delta}\left\{Z^*(\rho) \geq z_\alpha^*\right\} = 1 - \beta. \tag{15.13}$$

For applications that do not require dynamic response or covariate adaptive randomizations, it is often the practice to specify at design the randomization ratios $(r_0, r_1, \ldots, r_K)$ for which at least one ratio is exactly one. For example, let $r_0 = 1$. Then $n_k = n\, r_k$ for $k = 1, 2, \ldots, K$ where n is the sample size for the placebo group. With this simplification, the power on the left hand side of Eq. (15.13) only depends on n. Thus, the required n can be easily solved with a standard numerical univariate root-finding algorithm.

15.4 Likelihood Inference

15.4.1 Bootstrap Confidence Intervals

For any dose d such that $d \in (0, d_K]$, the discretized maximum likelihood estimate of $\Delta(d) = \Delta_{\max} f(d; \underline{\nu})$ is given by

$$\hat{\Delta}(d) = \hat{\Delta}_{\max} f(d; \hat{\underline{\nu}}) \tag{15.14}$$

following the three-step algorithm. However, the algorithm does not in itself provide estimates of standard errors for the estimates $\hat{\Delta}(d)$ for $d \in (0, d_K]$. The problem can be resolved via a bootstrap method. Specifically, we take parametric bootstrap samples from a multivariate normal distribution with mean $\hat{\underline{\Delta}}$ and variance-covariance matrix $\hat{\Sigma}$. By the three-step algorithm, these samples are first used to calculate the bootstrap distributions of the estimates $\tilde{\Delta}_{\max}$ and $\tilde{\underline{\nu}}$, which are in turn used to calculate the bootstrap estimates of $\tilde{\Delta}(d)$ for $d \in (0, d_K]$. Finally, we obtain the percentile bootstrap confidence intervals of $\hat{\Delta}(d)$ for $d \in (0, d_K]$ (Efron and Tibshirani 1993, p. 170).

15.4.2 Lower Effective Bound

For objective II, let $\mathcal{D} \subset (0, d_K]$ be the set of all potential doses of the drug formulation such that the dose response curve $\Delta(d) = \Delta_{\max} f(d; \nu)$ is increasing over $d \in \mathcal{D}$. Following Hemmings (2006, p. 47), we are interested in identifying all doses $d \in \mathcal{D}$ whose effect size $\Delta(d)$ are consistent with an effect size δ of clinical interest. The identification of such doses is achieved as follows.

For any $d \in \mathcal{D}$, let $\hat{\Delta}_L(d; \xi)$ be the lower bound of the $100(1 - 2\,\xi)\%$ confidence interval for $\Delta(d)$. Provided that the triple trends test is significant, i.e., $p^* \leq \alpha$ where p^* is given by Eq. (15.12), we would not eliminate dose d for future evaluation if the *test for statistical significance and clinical meaningfulness* (TSSCM) by

$$\hat{\Delta}_L(d; \xi) \geq \kappa_d,\ p^* \leq \alpha \tag{15.15}$$

is positive where $\kappa_d \geq 0$ is a superiority margin that satisfies

$$P_{\Delta(d)=\delta}\left\{\hat{\Delta}_L(d; \xi) \geq \kappa_d,\ p^* \leq \alpha\right\} \geq 1 - \beta \tag{15.16}$$

for which $P_{\Delta(d)=\delta}\{\cdot\}$ is the probability of a positive TSSCM under $\Delta(d) = \delta$ for dose d. The *lower effective bound* (LEB), $\hat{d}_{LE}(\alpha)$, is given by

$$\hat{d}_{LE}(\alpha) = \min\{d\colon \hat{\Delta}_L(d; \xi) \geq \kappa_d,\ p^* \leq \alpha \text{ for } d \in \mathcal{D}\}. \tag{15.17}$$

In case that it does not exist a dose $d \in \mathcal{D}$ such that $\hat{\Delta}_L(d;\xi) \geq \kappa_d$, $\hat{d}_{LE}(\alpha)$ could be set to a dose level, greater than d_K, that is consistent with the estimated dose response curve. For simplicity of calculation in simulation studies, we set $\hat{d}_{LE}(\alpha) = +\infty$.

The idea of the superiority margin κ_d for $d \in \mathcal{D}$ is based on a personal conversation of the author with Dr. Leber of the FDA's Division of Neurophamacological Drug Products in 1997. Dr. Leber suggested that there is no need for a statistical multiple testing procedure to identify a minimum effective dose as once a positive dose-response is established, the effect of any dose is also a positive; the issue is then how to identify a dose-range with a meaningful effect size of interest. A positive superiority margin κ_d for dose $d \in \mathcal{D}$ would prevent from choosing d if its effect size is very small. Following Eq. (15.16) the superiority margin κ_d for $d \in \mathcal{D}$ depends on the choice of δ. Thus, the lower effective bound $\hat{d}_{LE}(\alpha)$ defined in Eq. (15.17) is suitable for identifying doses with an observed treatment difference that is both statistically significant and consistent with a clinically meaningful effect size. This feature is fundamentally different from that of an estimated "minimum statistically significant dose" in the statistical literature whose determination only depends on meeting a statistical significance criterion. The breakthrough is that the lowest effect bound $\hat{d}_{LE}(\alpha)$ achieves a high probability of coverage for the threshold dose with clinically meaningful effect while completely avoids the ill-defined MED for dose-response models considered by the PhRMA working group.

15.4.3 Properties of LEB

For the given effect size δ, let

$$d_\delta = \min\{d\colon \Delta(d) \geq \delta \text{ and } d \in \mathcal{D}\} \tag{15.18}$$

be the minimum dose in $\mathcal{D}$. This implicitly assumes that $\Delta_{\max}$ is large enough that there is at least one dose d in $\mathcal{D}$ such that $\Delta(d) = \Delta_{\max} f(d;\underline{\nu}) \geq \delta$. Note that when $\Delta_{\max} < \delta$ there is no dose $d \in \mathcal{D}$ such that $\Delta(d) > \delta$. Thus, d_δ is not an intrinsic characteristics of the dose response curve $\Delta(d)$. For example, under the null hypothesis $H_0\colon \Delta_{\max} \leq 0$, d_δ does not exist.

To distinguish different dose response curves with respect to the effect size δ of interest, let $M_\delta(d_\delta;\underline{\nu})$ be an *effect-δ model* with dose response curve $\Delta(d)$ for which $\Delta(d) = \delta$ at $d = d_\delta \in \mathcal{D}$. The following theorem establishes the coverage probability property of the lower effective bound $\hat{d}_{LE}(\alpha)$ under an effect-δ model.

Theorem 1 *Let $\mathcal{D}_\delta = \{d\colon d \geq d_\delta, d \in \mathcal{D}\}$ for the effect-δ model $M_\delta(d_\delta;\underline{\nu})$ and $P_{M_\delta(d_\delta;\underline{\nu})}\{\cdot\}$ be the probability of an event under $M_\delta(d_\delta;\underline{\nu})$. Assume that*

$$P_{M_\delta(d_\delta;\underline{\nu})}\left[\cap_{d\in\mathcal{D}_\delta}\{\hat{\Delta}_L(d;\xi) \geq \kappa_d,\ p^* \leq \alpha\}\right] = 1-\beta. \tag{15.19}$$

Then

$$P_{M_\delta(d_\delta;\underline{\nu})}\left\{\hat{d}_{LE}(\alpha) \le d_\delta\right\} \ge 1-\beta. \tag{15.20}$$

The proof of the theorem is straightforward. Following the definition of $\hat{d}_{LE}(\alpha)$, it is clear that $\hat{\Delta}_L(d;\xi) \ge \kappa_d$ and $p^* \le \alpha$ for all $d \in \mathcal{D}_\delta$ implies that $\hat{d}_{LE}(\alpha) \le d_\delta$. Thus, the power requirement in Eq. (15.19) leads to the coverage property in Eq. (15.20).

The theorem links hypothesis testing by $\hat{\Delta}_L(d;\xi) \ge \kappa_d$ for $d \in \mathcal{D}_\delta$ to the identification of the lower effective bound $\hat{d}_{LE}(\alpha)$ for minimum dose d_δ with effect size δ. The probability requirement by Eq. (15.19) is motivated by the probability requirement for a classical screening objective, described by Bechhofer et al. (1995, pp. 126–130), for selecting all doses with treatment effects greater than the control. For our setting, *the screening objective*, or objective II can be more specifically stated as *to identify all potential doses with treatment effects greater than the placebo control by their superiority margins under the effect-δ models by* Eq. (15.19).

Another distinction is that the probability requirement by Eq. (15.19) applies to comparisons of any dose d in $\mathcal{D}$ which may or may not be a dose actually studies. That is d is not required to be from the set of doses $\{d_1, d_2, \ldots, d_K\}$ that is used in the study. In contrast, the probability requirement by Bechhofer et al. (1995, pp. 126–130) applies only to doses that are studied in the trial. While the notion of a minimum effective dose (MED) is widely used, it is surprising that this basic probability requirement is conspicuously absent from a very recent book on dose response trials by Ting (2006).

It is important to point out that the lower effective bound $\hat{d}_{LE}(\alpha)$ is not a confidence lower bound for a minimum dose d_δ. This is because under the effect-δ model $M_\delta(d_\delta;\underline{\nu})$, the coverage probability is specified by the power level $1-\beta$, rather than the confidence level $1-\alpha$. However, under a null model M_0 with $\Delta_{\max} = 0$ the coverage probability of $\hat{d}_{LE}(\alpha)$ for any $d \in D$ dramatically reduces to α. Formally, we have the following theorem.

Theorem 2

$$P_{M_0}\{\hat{d}_{LE}(\alpha) \le d\} \le \alpha \tag{15.21}$$

for all $d \in \mathcal{D}$.

By construction, $\hat{d}_{LE}(\alpha) \le d$ for any $d \in \mathcal{D}$ implies $p^* \le \alpha$. Thus, by Eq. (15.12), the null probability of coverage does not exceed α. We point out early that under the null hypothesis $H_0 : \Delta_{\max} \le 0$, d_δ does not exist. Therefore, for a given dose response trial, the probability of controlling the error rate of being able to identify any dose as d_δ needs to be controlled. Thus, the property given by Eq. (15.11) of Theorem 2 is important because under the null hypothesis H_0, the probability of falsely carrying any dose to phase 3 trials is controlled at the given α level.

Note that the same null probability requirement is used by Tamhane and Logan (2002). However, their MED does not incorporate the coverage probabilities, as in Eq. (15.20), under an effect-δ model for a given power $1-\beta$.

15.4.4 Criterion for Superiority Margins

The probability requirement by Eq. (15.19) in Theorem 1 is limited to a specific effect-δ model. In practice, there is little prior information to justify a particular effect-δ model, and therefore, the superiority margins κ_d for $d \in D$ may be required to be robust for all potential effect-δ models.

For each $d \in \mathcal{D}$, let there be an effect-δ model. The collect of all effect-δ models is denoted by $\mathcal{M}_\delta = \{M_\delta(d_\delta; \underline{\nu}) : d_\delta \in \mathcal{D}\}$. We choose the superiority margins κ_d for $d \in \mathcal{D}$ to satisfy the following robust power criterion:

$$\min_{M_\delta(d_\delta;\underline{\nu})\in\mathcal{M}_\delta} P_{M_\delta(d_\delta;\underline{\nu})}\left[\cap_{d\in\mathcal{D}_\delta}\{\hat{\Delta}_L(d;\xi) \geq \kappa_d,\, p^* \leq \alpha\}\right] \geq 1-\beta. \tag{15.22}$$

The determination of the superiority margins requires intensive numerical calculations and careful calibrations. The method is dependent upon the specific application in question. We illustrate principles for choosing the superiority margins for the case example in the following section.

15.5 Case Example

15.5.1 Trial Design

We reconsider the dose-response trial described by Schnitzer et al. (2004) to illustrate the proposed design. For simplicity, we only use 50 mg, 100 mg and 200 mg bid of lumiracoxib, in addition to a placebo control. Following the original trial design, we use 1:1:1:1 ratios to randomize patients to one of the three doses of lumiracoxib or placebo, and calculate the sample size to detect an 11-mm difference between the new drug and placebo for pain assessed on the VAS after 4 weeks of treatment, assuming a 20-mm standard deviation (SD) and a dropout rate of 10%. We use the one-sided significance level $\alpha = 0.025$ with a power of 95% (i.e., $\beta = 0.05$) to illustrate the efficiency of the triple trends trend. Following Sect. 15.3.4, we choose $d_{E_{90}} = 50$, 100, and 200 mg for the top, middle and bottom curve. The slope parameter ρ is set to be 1 and 3 for the top and middle curve. The required sample size is 77 patients per treatment group without adjusting for dropout. The slope parameter ρ for the middle curve is 2.35. With the specified 10% dropout rate, the adjusted per group sample size is 86, which is only one patient larger than 85 used for the original trial design for which the power is only 80%. The three sigmoidal $E_{\max}$ curves are shown in Fig. 15.1 in Sect. 15.3.4.

15.5.2 Calibration

In Hinz et al. (2009) it was demonstrated that a single dose 50 mg lumiracoxib inhibits blood monocyte COX-2 to similar degree as 100 or 200 mg. The question is raised whether doses lower than 100 mg are sufficient for pain therapy. Thus, it is possible that 25 mg bid is the "*less than fully effective dose that may still be useful*" (Temple 2004). To illustrate the proposed LEB approach, we assume that the results by Hinz et al. (2009) were available at the time the statistical analysis plan (SAP) was finalized. We would then be interested in inference on the efficacy of lumiracoxib at dose 25, 50, 75, 100, 150 and 200 mg bid. Notice that doses 25, 75 and 150 mg bid were not studies in the dose-response study.

The tree-step algorithm in Sect. 15.2.4 requires discretization of the shape vector $\underline{\nu} = (d_{E_{90}}, \rho)$. We choose $d_{E_{90}}$ ranging from 6.25 to 300 with the increment 6.25, and ρ from 0.5 to 10 with the increment 0.25. This leads to a 48-by-39 grid with a total of 1872 sigmoidal curves. The grid can be further expanded to include the very few special cases of the sigmoidal curves by Thomas (2006) for the subsets of *ad-hoc* models of by Bretz et al. (2005) that are not already covered by the 48-by-39 grid.

The most interesting as well as challenging aspect of the analysis planning is choosing the superiority margins κ_d for $d = 25, 50, 75, 100, 150$ and 200. There are several considerations.

(1) We want to limit the probability that 25 mg is chosen as the LEB if the effect size at dose 50 mg is 11-mm; however, the probability would need to be higher if the effect size at dose 50 mg is large, say, 15-mm.
(2) We require that the LEB has 95% (for $\beta = 0.05$) of probability coverage if the effect size at doses at or above 50 mg is 11-mm.
(3) We also want to limit the probabilities of choosing low doses (i.e., 50 or 75 mg) if the effect size 11-mm is achieved at higher doses.

To meet these requirements, we consider non-increasing superiority margin κ_d as a function of the dose $d \in \mathcal{D}$. We set $\kappa_{200} = 0$. Other values of the superiority margin κ_d are calibrated using the top curve for the triple trends test and a curve with the same shape function as the top curve but the effect size 15-mm at $d_1 = 50$ mg. Through trial-and-error with various configurations, we choose the one-sided lower confidence level 90% (i.e., $\xi = 0.1$) for the confidence lower bound $\hat{\Delta}_L(d; \xi)$ and obtain the superiority margins κ_d for $d = 25, 50, 75, 100, 150$ and 200 mg bid (see Table 15.3).

15.5.3 Simulation Studies

We use simulation studies to both calibrate the superiority margins and evaluate the operating characteristics of the dose-response analysis. This includes the distribution of the MED or "target-dose". Each simulation study has its own objective and

is conducted with a particular dose-response model. Following the trial design in Sect. 15.5.1, the sample size per group is 77 for all simulation studies.

We consider sigmoidal models $M_\delta(d_{E_{90}}, \rho)$ with threshold effect $\delta = 11$ at $d_{E_{90}}$. These effect-11 models are the three models used in the triple trends test and two additional models with $d_{E_{90}} = 75$ and 150, which are denoted by $M_{11}(50, 1)$, $M_{11}(75, 1.75)$, $M_{11}(100, 2.35)$, $M_{11}(150, 2.75)$ and $M_{11}(200, 3)$ in Table 15.1. The effect-11 models are chosen to represent the actual trial design of the case example for which the effect size $\delta = 11$ is used (Schnitzer et al. 2004). By the actual results of the trial, the models also reflect worst-case and yet practical settings that are not often used in the literature, with the exception of Kirby et al. (2011). We also consider the sigmoidal model $M_{15}(50, 1)$ to evaluate the probability of identifying dose

Table 15.1 Power, probability of coverage and distribution of estimated target dose

	Dose (mg b.i.d.)						
	25	50	75	100	150	200	∞
	$M_{11}(50, 1)$, Power = 0.9882						
ES	10	11	11.3793	11.5789	11.7857	11.8919	
PoCv LEB	0.4221	0.95	0.9654	0.9783	0.9882	0.9882	
Dstr MED	0.3036	0.2324	0.0745	0.0400	0.0241	0.0447	0.2806
	$M_{11}(75, 1.75)$, Power = 0.9853						
ES	6.9451	9.9700	11	11.4530	11.8313	11.9830	
PoCv LEB	0.3162	0.9301	0.9576	0.9734	0.9853	0.9853	
Dstr MED	0.2034	0.2625	0.1038	0.0596	0.0354	0.0555	0.2798
	$M_{11}(100, 2.35)$, Power = 0.9776						
ES	3.1436	7.8024	10.0309	11	11.7200	11.9615	
PoCv LEB	0.1519	0.8502	0.9410	0.9644	0.9776	0.9776	
Dstr MED	0.0632	0.2398	0.1596	0.1134	0.0615	0.0768	0.2857
	$M_{11}(150, 2.75)$, Power = 0.9654						
ES	1.1319	4.4736	7.5050	9.3572	11	11.5946	
PoCv LEB	0.0402	0.5375	0.8759	0.9375	0.9654	0.9654	
Dstr MED	0.0050	0.0766	0.1582	0.1865	0.1304	0.1130	0.3303
	$M_{11}(200, 3)$, Power = 0.9481						
ES	0.2111	1.5068	3.9338	6.4706	9.6743	11	
PoCv LEB	0.0082	0.1462	0.6811	0.8585	0.9481	0.9481	
Dstr MED	0.0003	0.0007	0.0395	0.1958	0.2156	0.1805	0.3615

ES—Effect size
PoCv LEB—Probability of coverage of lower effective bound (LEB)
Dstr MED—Distribution of MED

25 mg when its effect size is 15-mm. Both $M_{11}(50, 1)$ and $M_{15}(50, 1)$ are used for calibrating the superiority margins.

An important aspect of simulations studies is to evaluate the type 1 error rates of the LEB approach. The first and foremost is the numerical verification of Theorem 2, which addresses the type 1 error rate of falsely identifying a positive dose-response relationship and hence carrying forward any dose for phase 3 studies. The simulation study uses the null model M_0 with a flat dose-response curve. For a monotone dose-response model with a positive maximum effect size $\Delta_{\max}$, it is theoretically true that the effect size at any dose d is also positive. However, the magnitude of the effect size may be infinitesimal and thus controlling the multiple "*type 1 error rates*" for these doses is also important. We investigate this through the simulation study with models $M_7(150, 8)$ and $M_{11}(150, 8)$ such that the effect size for dose $d = 25$ and 50 mg bid is almost zero. Note that the shape parameter of the sigmoidal $E_{\max}$ model is $\rho = 8$, which is near the edge of the grid for ρ.

Occasionally, the observed dose-response curve can be bell-shaped, even though there is not biological reason that the reversal of effects are expected. This may be due to variability of the patients sample or imbalance of risk factors. To evaluate the performance of proposed methods, we also consider a biphasic dose-response model, denoted by M_{BiP}, whose parameters in Eq. (15.4) are $\delta = 11$ at $\eta = .9, \tau = 1$, $(D_1, \rho_1) = (100, 1.5)$, and $(D_2, \rho_2) = (200, 2)$.

A simulation study consists of bootstrapping nested within simulation runs. For each simulation run, a total of 10,000 bootstrap samples are drawn. The number of simulation runs is 11,000 for calibration and evaluation of power. The total of simulation runs of 33,000 are used for evaluating type 1 error rates under the null model. A supercomputer is used to perform all the simulation studies.

15.5.4 Results

Simulation results for effect-11 models are summarized in Table 15.1. The probabilities of coverage under model $M_{11}(50, 1)$ are not surprising as the superiority margins are calibrated with at the specified probability of coverage level 95%. The probabilities of coverage for the threshold dose $d_{E_{90}}$ under other models in Table 15.1 are all at or above 95%. This is in part due to the three superiority margins criteria in Sect. 15.5.2.

What is not known from the literature are the full distributions of the MED or "target-dose". The probability that the MED is right on-target is low; there is also a high probability that the MED or "target-dose" is either non-identifiable or above the largest study dose. This leads to the conclusion that the scheme of carrying over the MED or "target-dose" for phase 3 confirmative trials is flawed at a fundamental level; *phase 3 trials ought to not be used to confirm the efficacy or safety of an estimated dose with such an erratic distributional behavior*. Note that results on partial distributions of the MED or "target-dose" were obtained in the context of the PhRMA dose-ranging working group research and presented by the author at JSM in

Table 15.2 Special models

	Dose (mg b.i.d.)						
	25	50	75	100	150	200	∞
	M_0, Type I Error Rate = 0.0265						
ES	0	0	0	0	0	0	
PoCv LEB	0.0019	0.0107	0.0166	0.0199	0.0265	0.0265	
Dstr MED	0.0000	0.0000	0.0000	0.0001	0.0001	0.0019	0.9979
	$M_{15}(50, 1)$, Power = 0.9999						
ES	13.6364	15	15.5172	15.7895	16.0714	16.2162	
PoCv LEB	0.6565	0.9976	0.9987	0.9995	0.9999	0.9999	
Dstr MED	0.6345	0.3085	0.0238	0.0070	0.0037	0.0004	0.0186
	$M_7(150, 8)$, Power = 0.6620						
ES	0.0000	0.0107	0.2642	2.0214	7	7.6922	
PoCv LEB	0.0020	0.0180	0.1820	0.3720	0.6620	0.6620	
Dstr MED	0.0000	0.0000	0.0000	0.0160	0.0630	0.2470	0.6740
	$M_{11}(150, 8)$, Power = 0.9725						
ES	0.0000	0.0167	0.4151	3.1765	11	12.0878	
PoCv LEB	0.0012	0.0197	0.2873	0.7805	0.9725	0.9725	
Dstr MED	0.0000	0.0000	0.0036	0.1612	0.3528	0.3075	0.1745
	M_{BiP}, Power = 0.9853						
ES	3.8518	8.1432	10.8832	12.0712	11.5694	9.6085	
PoCv LEB	0.0885	0.9x21	0.9510	0.9556	0.9580	0.9580	
Dstr MED	0.0736	0.2827	0.1030	0.0327	0.0146	0.0235	0.4722

2006. However, the negative results were neither reported in Bornkamp et al. (2007) nor subsequently in Pinheiro et al. (2010).

Simulation results for the special models M_0, $M_{15}(50, 1)$, $M_7(150, 8)$, $M_{11}(150, 8)$ and M_{BiP} are summarized in Table 15.2. Under the null model M_0, the simulated type 1 error rate of the triple trends test is 0.0265, which is not statistically significant from the $\alpha = 0.025$ level. This assures that both the asymptotic theory as well as the implementation are correct. Thus, the LEB approach controls the overall type 1 error rate under model M_0 at the significance level $\alpha = 0.025$. For models $M_7(150, 8)$ and $M_{11}(150, 8)$, the multiple "type 1 error rates" for dose 25 and 50 mg bid are well below the $\alpha = 0.025$ level.

Under $M_{15}(50, 1)$, the LEB approach has a sight higher probability of picking up the 25 mg bid dose than the MED or "target-dose" scheme. This feature is achieved through the choice of superiority margins according to the calibration criteria in

Sect. 15.5.2. The probability of coverage 0.6565 does not affect the lowest dose 25 mg from being chosen: if the LEB = 25, then apparently 25 mg would be used for phase 3; if the LEB = 50, then 25 mg would be used again for phase 3 as the sub-effective dose of Temple (2004). On the other hand, if the probability of coverage is allowed to be much higher than 0.6565, then the probability of choosing no effect dose as the LEB, e.g., 25 or 50 mg under model $M_7(150, 8)$ or $M_{11}(150, 8)$, would be substantially higher. This is not a desirable property.

It is worthy of noting that under $M_7(150, 8)$ with $d_{E_{90}} = 7$, the maximum effect is $\Delta_{\max} = 7.7778$, which is less than $\delta = 11$. Thus, there is no dose under $M_7(150, 8)$ that would achieve the specified effect size $\delta = 11$. Yet, there is still a high chance (i.e., 32.6%) to falsely identify the "target-dose" at or below 200 mg.

The LEB approach performs well under the biphasic dose-response model M_{BiP}. The result is important to ensure that the LEB approach that relies solely on the sigmoidal $E_{\max}$ model is robust against "bell-shaped" occurrences in many clinical settings that may be due to the nature of the endpoint, imbalance of risk factors, or variability of the data, even though there is no biologic ground for reversal of efficacy at the high dose range. Note that if prior to or during the trial a possible biological ground for hormesis is postulated, modifications of the design and analysis on the basis of a biphasic dose-response model may be necessary. Of course, it is always possible to perform post-hoc analysis using a biphasic dose-response model when there is a justifiable cause.

It is observed that the probabilities of coverage for both 150 and 200 mg is exactly the power of the triple trends test for all monotone dose-response models. This is because that the superiority margins κ_{150} and κ_{200} are zero for both doses. However, whether a high dose (e.g., 150 or 200 mg) would be selected for phase 3 trials would depend on whether the dose is safe based on safety as well as laboratory data at hand. In the absence of these data, the decision would also depend on whether a high dose would provide additional efficacy benefit. With data that are consistent with model $M_7(150, 8)$, the decision to move forward to phase 3 with a high dose should also depend on if a subpopulation of patients could potentially have full efficacy benefits.

15.5.5 Probability of Success

The clinical development principles by Temple (2004) and Hemmings (2006, p. 47) can be numerically illustrated through the overall probability of technical, regulatory and post-approval success measure that incorporates a long-term benefit-risk profiles of the drug for a wider patient population. As a comparison, we also evaluate the probability of success of the "target-dose" scheme in the PhRMA working group reports (Bornkamp et al. 2007; Dragalin et al. 2010; Pinheiro et al. 2010).

For the top curve $M_{11}(50, 1)$, we assume the probabilities of success of 0.4, 0.8, 0.6, 0.4, 0.1 and 0 for carrying over the single dose of 25, 50, 75, 100, 150 and 200 mg for phase 3 study and marketing approval. Implicitly, the probability of success for failing to carry over a single dose for phase 3 is also zero. This scenario occurs

if statistical significance is not reached in phase 2 for a positive dose-response or the the "target-dose" is beyond 200 mg. For simplicity, the full dose-range is defined by the lower effective bound (LEB) and the highest dose 200 mg without a sub-effective dose. Then the probabilities of success for carrying over the all doses in the range [25, 200], [50, 200], [75, 200], [100, 200], [150, 200], and 200 are 0.8, 0.8, 0.6, 0.4, 0.1, and 0. Note that in practice, the probability of success with a full dose range could be much higher than the maximum of probabilities of success with individual doses.

With this particular setup, the probability of success of the development plan by Temple (2004) and Hemmings (2006, p. 47) is 77.54%, which is consistent with regulatory principles of clinical development. However, the probability of success of the "target-dose" scheme recommended by the PhRMA working group is only 37.05%. This exceedingly low probability of success is easily explained by the fact that the "target-dose" simply cannot be reliably identified or estimated when the threshold effect of 11-mm is already at the plateau of the dose-response (Table 15.1).

For the lumiracoxib case example, the choice of these probabilities are consistent with the development, post-marketing and regulatory history of lumiracoxib, as well as its totality of scientific and clinical data. In particular, the top curve is consistent with the results of the phase 2 dose-response study (Schnitzer et al. 2004) where the observed effect sizes for the physician global pain assessments are 10, 13.9, 13, 14.2 and 13.9-mm for lumiracoxib 50, 100, 200 mg bid, 400 mg od, and diclofenac 75 mg bid. The LEB approach is applied to the summary data with the actual reported sample size. The p-value for dose-response with the triple trends test is highly significant $1.4773e^{-7}$ and the LEB is 50 mg bid. The fitted curve as well as its confidence lower bound are given in Table 15.3. Thus, it suffices here to simply evaluate the overall probability of success of a particular development plan.

In prospective applications, however, different set of probabilities of success for other efficacy and safety scenarios are needed to ensure that the development plan is robust. This is illustrated for a neuropathic pain example in this author's 2006 JSM presentation, where similar results on identifiability and probability of success were

Table 15.3 Inference (fitted curve, confidence lower bound, and TSSCM)

d	Dose (mg b.i.d.)					
	25	50	75	100	150	200
$\hat{\delta}_d$		10		13.9		14.2
$\hat{\Delta}(d)$	0.0445	9.9698	13.3728	13.4572	13.4626	13.4627
$\hat{\Delta}_L(d;\xi)$	0.0332	6.0098	9.3426	9.9863	10.3311	10.4242
κ_d	0.1343	0.1343	0.1342	0.0166	0.0000	0.0000
TSSCM	0	1	1	1	1	1

$\hat{\delta}_d$—Observed effect size
$\hat{\Delta}(d)$—Fitted effect size
$\hat{\Delta}_L(d;\xi)$—Confidence lower bound for $\xi = 0.1$
κ_d—Superiority margin

observed for both the LEB approach and the "target-dose" scheme. However, the negative results of the "target-dose" scheme were absent from the PhRMA working group reports (Bornkamp et al. 2007; Pinheiro et al. 2010).

15.6 Discussion

15.6.1 Summary

We propose a test for statistical significance and clinical meaningfulness (TSSCM) for all feasible doses below a maximum dose studied in a phase 2 trial. A triple trends test is used to establish the overall statistical significance for a positive dose-response, while a likelihood test with superiority margins is used to detect an effect size consistent with a clinically meaningful difference. The superiority margins can be calibrated to meet a pre-specified power criterion of a threshold effect. The TSSCM is applied in a step-down fashion to identify the lower effective bound (LEB), which is then used to define the full dose-range for carrying forward to phase 3 trials. It is shown both theoretically and numerically via simulation studies that the LEB has desirable probabilities of coverage for the threshold dose of a positive dose-response relationship. The multiple type 1 error rates are also controlled. We stress that the LEB is not a confidence lower bound because its probability of coverage for the threshold dose can be maintained at any desirable level, say, 95%, under any effect-δ model (see Theorem 1) and the probability of coverage for any dose is controlled under the specified type 1 error rate α under the null model (see Theorem 2).

The LEB approach supports the regulatory principles for clinical development of Temple (2004) and Hemmings (2006, p. 47). We demonstrate that the idea of conducting full dose-range phase 3 studies after phase 2 dose-response trials is fundamentally sound through simulation evaluation of the overall probability of technical, regulatory and post-approval success that takes into account of the long-term benefit-risk profiles of approved dosing regimens for a wider patient population. In practice, this approach has lead to many successful pharmaceutical products with optimal benefit-risk profiles.

The likelihood test with superiority margins avoids the well-known difficulties associated estimating sigmoidal $E_{\max}$ model parameters such as $d_{E_{50}}$ or threshold dose $d_{E_{90}}$ (Sheiner et al. 1989). The core procedure for the likelihood test consists of calculating discretized maximum likelihood estimates of the maximum treatment effects with standard generalized linear model methods, selecting a best-fit-model over a grid of possible shape parameters, and bootstrapping the confidence intervals of treatment effects of individual doses. Not only is the core procedure simple to implement but also the derived likelihood based inference is robust against ill-behaved likelihood functions that make traditional likelihood estimates of the shape parameters difficult.

The three-step algorithm for fitting dose-response models applies to endpoints following a broad families of distributions as well as settings where the distributions cannot be fully specified. To assist trial design, we also provides a unified sample size procedure for the exponential family of distributions.

15.6.2 Phase 3 Full Dose-Range

Temple (2004) states that "*Having all or most phase 3 studies be D/R is usual for antihypertensives and antidepressants, anti-migraines, and anti-psychotics. This should be more common.*" We illustrate through the two case examples that Temple's full dose-range approach should also be extended to other disease areas (e.g., osteoarthritis and insomnia) in phase 3 clinical development. Questions still remain on the determination of the full dose-range and the choice of doses for the phase 3 trials.

The most interesting setting is when there are at least two doses that are identified to be effective from the phase 2 trial. The easiest dose is the LEB. A dose lower than the LEB, which is expected to be less than fully effective, should also be included. Following Temple (2004) the objective is to study "*whether sub-effective dose represents some people responding fully or all people responding a little.*" Such information provide a critical role in drug labeling, which is especially important as phase 3 trials are often conducted with longer duration of treatment and follow-up in a more heterogeneous patient population. A dose higher than the LEB should be based on the evaluation of efficacy and short term safety as well as laboratory values. Other information from PK/PD analysis, or Biomarker/PD analysis may also be useful for dose selection.

For the case example with lumiracoxib, the analysis of the phase 2 data identified 50 mg bid as the LEB. Following Sect. 15.5.2, the lower dose of 25 mg big may be chosen based on the results by Hinz et al. (2009). There are no dose-response related adverse events at or above 100 mg bid. As doses higher than 100 mg bid do not provide additional efficacy benefits, 100 mg bid should be used as the high dose.

A similar analysis with suvorexant is also performed, by which 10 mg is the LEB. The low dose 5 mg may be the less than the fully effective dose. Based on evaluations of available phase 2 safety and laboratory values, the high dose 20 mg should be chosen. This full dose-range is identical to the requirement in the FDA's complete letter. However, this decision could be reached much early, i.e., prior to initiation of the suvorexant phase 3 program.

15.6.3 Future Work

The paper addresses issues on design and inference for phase 2 trials with the focus to determine the full dose-range and doses for the phase 3 clinical development

program. While the structure of the phase 3 program may vary according the the disease area or the nature of the new drug, there should be a least one phase 3 trial with a full dose-range. The methods developed here in this paper can extended to such phase 3 trial as well as to meta-analysis of all phase 3 trials. For individual phase 3 dose-response trials, the methods reduce the complexity of multiplicity due to multiple doses. In addition to being able to identify the lower effective bound (LEB), additional work is necessary to develop methods for defining the upper safety bound (USB) with an acceptable safety threshold. Both LEB and USB can then be used to describe the therapeutic window for the drug. Additional research is also necessary to address Temple's question of "*whether sub-effective dose represents some people responding fully or all people responding a little.*" We note that the ability to address this question can be reduced if the phase 3 trial would drop the low dose through futility analysis. This raises concerns of many "adaptive designs" that may already be in use.

The paper is limited to settings where parallel group designs may be the only choice. There are other settings where cross-over designs are often used. However, for clinical endpoints, rather than pharmacodynamic (PD) endpoints where the effects are often tied to the drug concentration, a typical washout window may not prevent the presence of carry-over effects. This often complicates the analysis and its interpretation (e.g., suvorexant). As a result, analysis and interpretation are then restricted to data from the first period, which is a parallel group design. In addition, cross-over designs do not handle informative dropouts, which are common in clinical trials. The essence of a cross-over design for most clinical settings with possible carry-over effects and dropouts are to enhance the efficacy of the trial with multiple use of the same patients so as to reduce the number of patients enrolled. Research is needed to develop new dose-response designs to achieve this objective while avoiding the problem of carry-over effects and lessening the impacts of informative dropouts.

The proposed dose-response designs need to be expanded to settings where hormesis is justified on biological ground. The key is to work with the biphasic dose-response model proposed in this paper. Issues relating the number of doses, sample size calculation and test for dose-response relationship need to incorporate more models that are carefully chosen. The grid for discretized likelihood method also need to be carefully constructed.

Future work may also consider the setting where an active control is used for assessing assay sensitivity and evaluation of *comparative effectiveness*, which may be similarly defined through *lower comparative effective bound* (LCEB). We also envision a broader used of TSSCM for clinical trial designs in general. A nature setting is non-inferiority trials where a statistical non-inferiority margin is used for inference that a new drug is effective against a putative placebo while a clinical non-inferiority margin is used to establish that the new drug is similar to an active control.

Last but not the least, the analytics and supercomputer used for this paper may be accessible through collaborative research and development for new research projects. Contingent upon availability of future funding, the author intends to deploy the analytics and supercomputing through a Software as a Service (SaaS) cloud infrastructure.

Acknowledgements The authors like to thank two reviewers for their constructive comments and suggestions.

References

Bechhofer, R. E., Santner, T. J., & Goldsman, D. M. (1995). *Design and analysis of experiments for statistical selection, screening, and multiple comparisons*. New York: Wiley.

Bornkamp, B., Frank Bretz, F., et al. (2007). Innovative approaches for designing and analyzing adaptive dose-ranging trials. *Journal of Biopharmaceutical Statistics*, *17*, 965–995.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Bretz, F., Pinheiro, J., & Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, *61*, 738–748.

Capizzi, T., Survill, T. T., Heyse, J. F., & Malani, H. (1992). An empirical and simulated comparison of some tests for detecting progressiveness of response with increasing doses of a compound. *Biometrical Journal*, *34*, 275–289.

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. London: Chapman & Hall.

Dragalin, V., Bornkamp, B., et al. (2010). A simulation study to compare new adaptive dose-ranging designs. *Statistics in Biopharmaceutical Research*, *2*, 487–512.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

Fleischmann, R., Hyman Tannenbaum, H., et al. (2008). Long-term retention on treatment with lumiracoxib 100 mg once or twice daily compared with celecoxib 200 mg once daily: A randomised controlled trial in patients with osteoarthritis. *BMC Musculoskeletal Disorders*, *9*, 1–32.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, *1*, 141–149.

Hemmings, R. (2006). Philosophy and methodology of dose-finding a regulatory perspective. In S. Chevret (Ed.), *Statistical methods for dose-finding experiments*. New York: Wiley.

Hinz, B., Renner, B., et al. (2009). Lumiracoxib inhibits cyclooxygenase 2 completely at the 50 mg dose: is liver toxicity avoidable by adequate dosing? *Annals of the Rheumatic Diseases*, *68*, 289–291.

Javaherian, K., Lee, T. Y., et al. (2011). Two endogenous antiangiogenic inhibitors, endostatin and angiostatin, demonstrate biphasic curve in their antitumor profiles. *Dose-Response*, *9*, 369–376.

Kabbinavar, F., Hurwitz, H. I., et al. (2003). Phase II, randomized trial comparing bevacizumab plus fluorouracil (FU)/leucovorin (LV) with FU/LV alone in patients with metastatic colorectal cancer. *Journal of Clinical Oncology*, *21*, 60–65.

Kirby, S., Brain, P., & Jones, B. (2011). Fitting $E_{\max}$ models to clinical trial dose response data. *Pharmaceutical Statistics*, *10*, 143–149.

Lalonde, R. L. (1992). Pharmacodynamics. In W. E. Evans, J. J. Schentag, & W. J. Jusko (Eds.), *Applied pharmacokinetics: Principles of therapeutic drug monitoring* (3rd ed.). Vancouver, WA: Applied Therapeutics Inc.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Liao, Z., & Liu, R. (2009). Re-parameterization of five-parameter logistic function. *Journal of Chemometrics*, *23*, 248–253.

Liu, Q., & Pierce, D. A. (1993). Heterogeneity in Mantel-Haenszel-type models. *Biometrika*, *80*, 543–556.

Liu, Q., & Pledger, W. G. (2005). Phase 2 and 3 combination designs to accelerate drug development. *Journal of the American Statistical Association*, *100*, 493–502.

Lockwood, P. A., Cook, J. A., Ewy, W. E., & Mandema, J. W. (2003). The use of clinical trial simulation to support dose selection: Application to development of a new treatment for chronic neuropathic pain. *Pharmaceutical Research*, *20*, 1752–1759.

Parker, I. (2013). The Big Sleep: Insomnia drugs like Ambien are notorious for their side effects. Has Merck created a blockbuster replacement? *The New Yorker*.

Pinheiro, J., Sax, F., et al. (2010). Adaptive and model-based dose-ranging trials: Quantitative evaluation and recommendations. White paper of the PhRMA Working Group on adaptive dose-ranging studies. *Statistics in Biopharmaceutical Research*, *2*, 435–454.

Quan, H., & Capizzi, T. (1999). Adjusted regression trend test for a multicenter clinical trial. *Biometrics*, *55*, 460–462.

Reynolds, A. R. (2010). Potential relevance of bell-shaped and u-shaped dose-responses for the therapeutic targeting of angiogenesis in cancer. *Dose-Response*, *8*, 253–284.

Ruberg, S. J. (1989). Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association*, *84*, 816–822.

Schnitzer, T. J., Burmester, G. R., et al. (2004). Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), reduction in ulcer complications: Randomised controlled trial. *Lancet*, *364*, 665–674.

Sheiner, L. B., Beal, S. L., & Sambol, N. C. (1989). Study designs for dose-ranging. *Clinical Pharmacology & Therapeutics*, *46*, 63–77.

Sheiner, L. B., Hashimoto, Y., & Beal, S. L. (1991). A simulation study comparing designs for dose ranging. *Statistics in Medicine*, *10*, 303–21.

Sheldon, E. A., Beaulieu, A., et al. (2008). Long-term efficacy and safety of lumiracoxib 100 mg: An open-label extension of a 13-week randomized controlled trial in patients with primary osteoarthritis of the knee. *Clinical and Experimental Rheumatology*, *26*, 611–619.

Tamhane, A. C., & Logan, B. R. (2002). Multiple test procedures for identifying the minimum effective and maximum safe doses of a drug. *Journal of the American Statistical Association*, *97*, 293–301.

Temple, R. (2004). The Critical path opportunities for efficiency in development. In *FDA Science Board Advisory Committee Meeting*.

Thomas, N. (2006). Hypothesis testing and Bayesian estimation using a sigmoid $E_{\max}$ model applied to sparse dose designs. *Journal of Biopharmaceutical Statistics*, *16*, 657–677.

Ting, N. (2006). *Dose finding in drug development*. New York: Spinger.

Tukey, J. W., Ciminera, J. L., & Heyse, J. F. (1985). Testing the statistical certainty of a response to increasing dose of a drug. *Biometrics*, *41*, 295–301.

Wagner, J. G. (1968). Kinetics of pharmacologic response. *Journal of Theoretical Biology*, *20*, 173–201.

Index

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium*, ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7829-3