



# Analysis of Clinical Trials Using SAS®

A Practical Guide

*Second Edition*

Edited by  
Alex Dmitrienko  
Gary G. Koch

The correct bibliographic citation for this manual is as follows: Dmitrienko, Alex, and Gary G. Koch. 2017. *Analysis of Clinical Trials Using SAS®: A Practical Guide, Second Edition*. Cary, NC: SAS Institute Inc.

**Analysis of Clinical Trials Using SAS®: A Practical Guide, Second Edition**

Copyright © 2017, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-62959-847-5 (Hard copy)

ISBN 978-1-63526-144-8 (EPUB)

ISBN 978-1-63526-145-5 (MOBI)

ISBN 978-1-63526-146-2 (PDF)

All Rights Reserved. Produced in the United States of America.

**For a hard-copy book:** No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

**For a web download or e-book:** Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

**U.S. Government License Rights; Restricted Rights:** The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

July 2017

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

# Contents

<b>Preface</b>	v
<b>About This Book</b>	xi
<b>About These Authors</b>	xii
<b>1 Model-based and Randomization-based Methods</b>	1
<b>By Alex Dmitrienko and Gary G. Koch</b>	
1.1 Introduction	1
1.2 Analysis of continuous endpoints	4
1.3 Analysis of categorical endpoints	20
1.4 Analysis of time-to-event endpoints	41
1.5 Qualitative interaction tests	56
References	61
<b>2 Advanced Randomization-based Methods</b>	67
<b>By Richard C. Zink, Gary G. Koch, Yunro Chung and Laura Elizabeth Wiener</b>	
2.1 Introduction	67
2.2 Case studies	70
2.3 %NParCov4 macro	73
2.4 Analysis of ordinal endpoints using a linear model	74
2.5 Analysis of binary endpoints	78
2.6 Analysis of ordinal endpoints using a proportional odds model	79
2.7 Analysis of continuous endpoints using the log-ratio of two means	80
2.8 Analysis of count endpoints using log-incidence density ratios	81
2.9 Analysis of time-to-event endpoints	82
2.10 Summary	86
<b>3 Dose-Escalation Methods</b>	101
<b>By Guochen Song, Zoe Zhang, Nolan Wages, Anastasia Ivanova, Olga Marchenko and Alex Dmitrienko</b>	
3.1 Introduction	101
3.2 Rule-based methods	103
3.3 Continual reassessment method	107
3.4 Partial order continual reassessment method	116
3.5 Summary	123
References	123

<b>4 Dose-finding Methods</b>	<b>127</b>
<b>By Srinand Nandakumar, Alex Dmitrienko and Ilya Lipkovich</b>	
4.1 Introduction	127
4.2 Case studies	128
4.3 Dose-response assessment and dose-finding methods	132
4.4 Dose finding in Case study 1	145
4.5 Dose finding in Case study 2	160
References	176
<b>5 Multiplicity Adjustment Methods</b>	<b>179</b>
<b>By Thomas Brechenmacher and Alex Dmitrienko</b>	
5.1 Introduction	179
5.2 Single-step procedures	184
5.3 Procedures with a data-driven hypothesis ordering	189
5.4 Procedures with a prespecified hypothesis ordering	202
5.5 Parametric procedures	212
5.6 Gatekeeping procedures	221
References	241
Appendix	244
<b>6 Interim Data Monitoring</b>	<b>251</b>
<b>By Alex Dmitrienko and Yang Yuan</b>	
6.1 Introduction	251
6.2 Repeated significance tests	253
6.3 Stochastic curtailment tests	292
References	315
<b>7 Analysis of Incomplete Data</b>	<b>319</b>
<b>By Geert Molenberghs and Michael G. Kenward</b>	
7.1 Introduction	319
7.2 Case Study	322
7.3 Data Setting and Methodology	324
7.4 Simple Methods and MCAR	334
7.5 Ignorable Likelihood (Direct Likelihood)	338
7.6 Direct Bayesian Analysis (Ignorable Bayesian Analysis)	341
7.7 Weighted Generalized Estimating Equations	344
7.8 Multiple Imputation	349
7.9 An Overview of Sensitivity Analysis	362
7.10 Sensitivity Analysis Using Local Influence	363
7.11 Sensitivity Analysis Based on Multiple Imputation and Pattern-Mixture Models	371
7.12 Concluding Remarks	378
References	378
<b>Index</b>	<b>385</b>

# Preface

## Introduction

---

Clinical trials have long been one of the most important tools in the arsenal of clinicians and scientists who help develop pharmaceuticals, biologics, and medical devices. It is reported that close to 10,000 clinical studies are conducted every year around the world. We can find many excellent books that address fundamental statistical and general scientific principles underlying the design and analysis of clinical trials. [for example, Pocock (1983); Fleiss (1986); Meinert (1986); Friedman, Furberg, and DeMets (1996); Piantadosi (1997); and Senn (1997)]. Numerous references can be found in these fine books. It is also important to mention recently published SAS Press books that discuss topics related to clinical trial statistics as well as other relevant topics, e.g., Dmitrienko, Chuang-Stein, and D'Agostino (2007); Westfall, Tobias, and Wolfinger (2011); Stokes, Davis, and Koch (2012); and Menon and Zink (2016).

The aim of this book is unique in that it focuses in great detail on a set of selected and practical problems facing statisticians and biomedical scientists conducting clinical research. We discuss solutions to these problems based on modern statistical methods, and we review computer-intensive techniques that help clinical researchers efficiently and rapidly implement these methods in the powerful SAS environment.

It is a challenge to select a few topics that are most important and relevant to the design and analysis of clinical trials. Our choice of topics for this book was guided by the International Conference on Harmonization (ICH) guideline for the pharmaceutical industry entitled "Structure and Content of Clinical Study Reports," which is commonly referred to as ICH E3. The documents states the following:

"Important features of the analysis, including the particular methods used, adjustments made for demographic or baseline measurements or concomitant therapy, handling of dropouts and missing data, adjustments for multiple comparisons, special analyses of multicenter studies, and adjustments for interim analyses, should be discussed [in the study report]."

Following the ICH recommendations, we decided to focus in this book on the analysis of stratified data, incomplete data, multiple inferences, and issues arising in safety and efficacy monitoring. We also address other statistical problems that are very important in a clinical trial setting. The latter includes reference intervals for safety and diagnostic measurements.

One special feature of the book is the inclusion of numerous SAS macros to help readers implement the new methodology in the SAS environment. The availability of the programs and the detailed discussion of the output from the macros help make the applications of new procedures a reality.

The book is aimed at clinical statisticians and other scientists who are involved in the design and analysis of clinical trials conducted by the pharmaceutical industry and academic institutions or governmental institutions, such as NIH. Graduate

students specializing in biostatistics will also find the material in this book useful because of the applied nature of this book.

Since the book is written for practitioners, it concentrates primarily on solutions rather than the underlying theory. Although most of the chapters include some tutorial material, this book is not intended to provide a comprehensive coverage of the selected topics. Nevertheless, each chapter gives a high-level description of the methodological aspects of the statistical problem at hand and includes references to publications that contain more advanced material. In addition, each chapter gives a detailed overview of the key statistical principles. References to relevant regulatory guidance documents, including recently released guidelines on adaptive designs and multiplicity issues in clinical trials, are provided. Examples from multiple clinical trials at different stages of drug development are used throughout the book to motivate and illustrate the statistical methods presented in the book.

## Outline of the book

---

The book has been reorganized based on the feedback provided by numerous readers of the first edition. The topics covered in the second edition are grouped into three parts. The first part (Chapters 1 and 2) provides detailed coverage of general statistical methods used at all stages of drug development. Further, the second part (Chapters 3 and 4) and third part (Chapters 5, 6, and 7) focus on the topics specific to early-phase and late-phase clinical trials, respectively.

The chapters from the first edition have been expanded to cover new approaches to addressing the statistical problems introduced in the original book. Numerous revisions have been made to improve the explanations of key concepts and to add more examples and case studies. A detailed discussion of new features of SAS procedures has been provided. In some cases, new procedures are introduced that were not available when the first edition was released.

A brief outline of each chapter is provided below. New topics are carefully described and expanded coverage of the material from the first edition is highlighted.

### Part I: General topics

As stated above, the book opens with a review of a general class of statistical methods used in the analysis of clinical trial data. This includes model-based and non-parametric approaches to examining the treatment effect on continuous, categorical, count, and time-to-event endpoints. Chapter 1 is mostly based on a chapter from the first edition. Chapter 2 has been added to introduce versatile randomization-based methods for estimating covariate-adjusted treatment effects.

#### **Chapter 1 (Model-based and Randomization-based Methods)**

Adjustments for important covariates such as patient baseline characteristics play a key role in the analysis of clinical trial data. The goal of an adjusted analysis is to provide an overall test of the treatment effect in the presence of prognostic factors that influence the outcome variables of interest. This chapter introduces model-based and non-parametric randomization-based methods commonly used in clinical trials with continuous, categorical, and time-to-event endpoints. It is assumed that the covariates of interest are nominal or ordinal. Thus, they can be used to define strata, which leads to a stratified analysis of relevant endpoints. SAS implementation of these statistical methods relies on PROC GLM, PROC FREQ, PROC LOGISTIC, PROC GENMOD, and other procedures. In addition, the chapter introduces statistical

methods for studying the nature of treatment-by-stratum interactions. Interaction tests are commonly carried out in the context of subgroup assessments. A popular treatment-by-stratum interaction test is implemented using a custom macro.

## **Chapter 2 (Advanced Randomization-based Methods)**

This chapter presents advanced randomization-based methods used in the analysis of clinical endpoints. This class of statistical methods complements traditional model-based approaches. In fact, clinical trial statisticians are encouraged to consider both classes of methods since each class is useful within a particular setting, and the advantages of each class offset the limitations of the other class. The randomization-based methodology relies on minimal assumptions and offers several attractive features, e.g., it easily accommodates stratification and supports essentially-exact  $p$ -values and confidence intervals. Applications of advanced randomization-based methods to clinical trials with continuous, categorical, count, and time-to-event endpoints are presented in the chapter. Randomization-based methods are implemented using a powerful SAS macro (%NParCov4) that is applicable to a variety of clinical outcomes.

## **Part II: Early-phase clinical trials**

Chapters 3 and 4 focus on statistical methods that commonly arise in Phase I and Phase II trials. These chapters are new to the second edition and feature a detailed discussion of designs used in dose-finding trials, dose-response modeling, and identification of target doses.

### **Chapter 3 (Dose-Escalation Methods)**

Dose-ranging and dose-finding trials are conducted at early stages of all drug development programs to evaluate the safety and often efficacy of experimental treatments. This chapter gives an overview of dose-finding methods used in dose-escalation trials with emphasis on oncology trials. It provides a review of basic dose-escalation designs, and focuses on powerful model-based methods such as the continual reassessment method for trials with a single agent and its extension (partial order continual reassessment method) for trials with drug combinations. Practical issues related to the implementation of model-based methods are discussed and illustrated using examples from Phase I oncology trials. Custom macros that implement the popular dose-finding methods used in dose-escalation trials are introduced in this chapter.

### **Chapter 4 (Dose-Finding Methods)**

Identification of target doses to be examined in subsequent Phase III trials plays a central role in Phase II trials. This new chapter introduces a class of statistical methods aimed at examining the relationship between the dose of an experimental treatment and clinical response. Commonly used approaches to testing dose-response trends, estimating the underlying dose-response function, and identifying a range of doses for confirmatory trials are presented. Powerful contrast-based methods for detecting dose-response signals evaluate the evidence of treatment benefit across the trial arms. These methods emphasize hypothesis testing. But they can be extended to hybrid methods that combine dose-response testing and dose-response modeling to provide a comprehensive approach to dose-response analysis (MCP-Mod procedure). Important issues arising in dose-response modeling, such as covariate adjustments and handling of missing observations, are discussed in the chapter. Dose-finding methods discussed in the chapter are implemented using SAS procedures and custom macros.

## Part III: Late-phase clinical trials

The following three chapters focus on statistical methods commonly used in late-phase clinical trials, including confirmatory Phase III trials. These chapters were included in the first edition of the book. But they have undergone substantial revisions to introduce recently developed statistical methods and to describe new SAS procedures.

### **Chapter 5 (Multiplicity Adjustment Methods)**

Multiplicity arises in virtually all late-phase clinical trials—especially in confirmatory trials that are conducted to study the effect of multiple doses of a novel treatment on several endpoints or in several patient populations. When multiple clinical objectives are pursued in a trial, it is critical to evaluate the impact of objective-specific decision rules on the overall Type I error rate. Numerous adjustment methods, known as multiple testing procedures, have been developed to address multiplicity issues in clinical trials. The revised chapter introduces a useful classification of multiple testing procedures that helps compare and contrast candidate procedures in specific multiplicity problems. A comprehensive review of popular multiple testing procedures is provided in the chapter. Relevant practical considerations and issues related to SAS implementation based on SAS procedures and custom macros are discussed. A detailed description of advanced multiplicity adjustment methods that have been developed over the past 10 years, including gatekeeping procedures, has been added in the revised chapter. A new macro (`%MixGate`) has been introduced to support gatekeeping procedures that have found numerous applications in confirmatory clinical trials.

### **Chapter 6 (Interim Data Monitoring)**

The general topic of clinical trials with data-driven decision rules, known as adaptive trials, has attracted much attention across the clinical trial community over the past 15-20 years. This chapter uses a tutorial-style approach to introduce the most commonly used class of adaptive trial designs, namely, group-sequential designs. It begins with a review of repeated significance tests that are broadly applied to define decision rules in trials with interim looks. The process of designing group sequential trials and flexible procedures for monitoring clinical trial data are described using multiple case studies. In addition, the chapter provides a survey of popular approaches to setting up futility tests in clinical trials with interim assessments. These approaches are based on frequentist (conditional power), mixed Bayesian-frequentist (predictive power), and fully Bayesian (predictive probability) methods. The updated chapter takes advantage of powerful SAS procedures (PROC SEQDESIGN and PROC SEQTEST) that support a broad class of group-sequential designs used in clinical trials.

### **Chapter 7 (Analysis of Incomplete Data)**

A large number of empirical studies are prone to incompleteness. Over the last few decades, a number of methods have been developed to handle incomplete data. Many of those are relatively simple, but their performance and validity remain unclear. With increasing computational power and software tools available, more flexible methods have come within reach. The chapter sets off by giving an overview of simple methods for dealing with incomplete data in clinical trials. It then focuses on ignorable likelihood and Bayesian analyses, as well as on weighted generalized estimating equations (GEE). The chapter considers in detail sensitivity analysis tools to explore the impact that not fully verifiable assumptions about the missing data mechanism have on ensuing inferences. The original chapter has been extended

by including a detailed discussion of PROC GEE with emphasis on how it can be used to conduct various forms of weighted generalized estimating equations analyses. For sensitivity analysis, the use of the MNAR statement in PROC MI is given extensive consideration. It allows clinical trial statisticians to vary missing data assumptions, away from the conventional MAR (missing at random) assumption.

## About the contributors

---

This book has been the result of a collaborative effort of 16 statisticians from the pharmaceutical industry and academia:

**Thomas Brechenmacher**, Statistical Scientist, Biostatistics, QuintilesIMS.

**Yunro Chung**, Postdoctoral Research Fellow, Public Health Sciences Division, Fred Hutchinson Cancer Research Center.

**Alex Dmitrienko**, President, Mediana Inc.

**Anastasia Ivanova**, Associate Professor of Biostatistics, University of North Carolina at Chapel Hill.

**Michael G. Kenward**, Professor of Biostatistics, Luton, United Kingdom.

**Gary G. Koch**, Professor of Biostatistics and Director of the Biometrics Consulting Laboratory at the University of North Carolina at Chapel Hill.

**Ilya Lipkovich**, Principal Scientific Advisor, Advisory Analytics, QuintilesIMS.

**Olga Marchenko**, Vice President, Advisory Analytics, QuintilesIMS.

**Geert Molenberghs**, Professor of Biostatistics, I-BioStat, Universiteit Hasselt and KU Leuven, Belgium.

**Srinand Nandakumar**, Manager of Biostatistics, Global Product Development, Pfizer.

**Guochen Song**, Associate Director, Biostatistics, Biogen.

**Nolan Wages**, Assistant Professor, Division of Translational Research and Applied Statistics, Department of Public Health Sciences, University of Virginia.

**Laura Elizabeth Wiener**, Graduate Student, University of North Carolina at Chapel Hill.

**Yang Yuan**, Distinguished Research Statistician Developer, SAS Institute Inc.

**Zoe Zhang**, Statistical Scientist, Biometrics, Genentech.

**Richard C. Zink**, Principal Research Statistician Developer, JMP Life Sciences, SAS Institute Inc., and Adjunct Assistant Professor, University of North Carolina at Chapel Hill.

## Acknowledgments

---

We would like to thank the following individuals for a careful review of the individual chapters in this book and valuable comments (listed in alphabetical order): Brian Barkley (University of North Carolina at Chapel Hill), Emily V. Dressler (University of Kentucky), Ilya Lipkovich (QuintilesIMS), Gautier Paux (Institut de Recherches Internationales Servier), and Richard C. Zink (JMP Life Sciences, SAS Institute).

We are grateful to Brenna Leath, our editor at SAS Press, for her support and assistance in preparing this book.

## References

---

- Dmitrienko, A., Chuang-Stein, C., D'Agostino, R. (editors) (2007). *Pharmaceutical Statistics Using SAS*. Cary, NC: SAS Institute, Inc.
- Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: John Wiley.
- Friedman, L.M., Furberg, C.D., DeMets, D.L. (1996). *Fundamentals of Clinical Trials*. St. Louis, MO: Mosby-Year Book.
- Meinert, C.L. (1986). *Clinical Trials: Design, Conduct and Analysis*. New York: Oxford University Press.
- Menon, S., Zink, R. (2016). *Clinical Trials Using SAS: Classical, Adaptive and Bayesian Methods*. Cary, NC: SAS Press.
- Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*. New York: John Wiley.
- Pocock, S.J. (1983). *Clinical Trials: A Practical Approach*. New York: John Wiley.
- Senn, S.J. (2008). *Statistical Issues in Drug Development*. Second Edition. Chichester: John Wiley.
- Stokes, M., Davis, C.S., Koch, G.G. (2012). *Categorical Data Analysis Using SAS*. Third Edition. Cary, NC: SAS Press.
- Westfall, P.H., Tobias, R.D., Wolfinger, R.D. (2011). *Multiple Comparisons and Multiple Tests Using SAS*. Second Edition. Cary, NC: SAS Institute, Inc.

# About This Book

---

## What Does This Book Cover?

The main goal of this book is to introduce popular statistical methods used in clinical trials and to discuss their implementation using SAS software. To help bridge the gap between modern statistical methodology and clinical trial applications, the book includes numerous case studies based on real trials at all stages of drug development. It also provides a detailed discussion of practical considerations and relevant regulatory issues as well as advice from clinical trial experts.

The book focuses on fundamental problems arising in the context of clinical trials such as the analysis of common types of clinical endpoints and statistical approaches most commonly used in early- and late-stage clinical trials. The book provides detailed coverage of approaches utilized in Phase I/Phase II trials, e.g., dose-escalation and dose-finding methods. Important trial designs and analysis strategies employed in Phase II/Phase III include multiplicity adjustment methods, data monitoring methods and methods for handling incomplete data.

---

## Is This Book for You?

Although the book was written primarily for biostatisticians, the book includes high-level introductory material that will be useful for a broad group of pre-clinical and clinical trial researchers, e.g., drug discovery scientists, medical scientists and regulatory scientists working in the pharmaceutical and biotechnology industries.

---

## What Are the Prerequisites for This Book?

General experience with clinical trials and drug development, as well as experience with SAS/STAT procedures, will be desirable.

---

## What's New in This Edition?

The second edition of this book has been thoroughly revised based on the feedback provided by numerous readers of the first edition. The topics covered in the book have been grouped into three parts. The first part provides detailed coverage of general statistical methods used across the three stages of drug development. The second and third parts focus on the topics specific to early-phase and late-phase clinical trials, respectively.

The chapters from the first edition have been expanded to cover new approaches to addressing the statistical problems introduced in the original book. Numerous revisions have been made to improve the explanations of key concepts, add more examples and case studies. A detailed discussion of new features of SAS procedures has been provided and, in some cases, new procedures are introduced that were not available when the first edition was released.

---

## What Should You Know about the Examples?

The individual chapters within this book include tutorial material along with multiple examples to help the reader gain hands-on experience with SAS/STAT procedures used in the analysis of clinical trials.

---

### Software Used to Develop the Book's Content

The statistical methods introduced in this book are illustrated using numerous SAS/STAT procedures, including PROC GLM, PROC FREQ, PROC LOGISTIC, PROC GENMOD, PROC LIFETEST and PROC PHREG (used in the analysis of different types of clinical endpoints), PROC MIXED, PROC NLMIXED and PROC GENMOD (used in dose-finding trials), PROC MULTTEST (used in clinical trials with multiple objectives), PROC SEQDESIGN and PROC SEQTEST (used in group-sequential trials), PROC MIXED, PROC GLIMMIX, PROC GEE, PROC MI and PROC MIANALYZE (used in clinical trials with missing data). These procedures are complemented by multiple SAS macros written by the chapter authors to support advanced statistical methods.

---

### Example Code and Data

You can access the example code, SAS macros and data sets used in this book by linking to its author page at <http://support.sas.com/publishing/authors/dmitrienko.html>.

---

### SAS University Edition

 This book is compatible with SAS University Edition. If you are using SAS University Edition, then begin here: <https://support.sas.com/ue-data>.

---

### Output and Graphics

The second edition takes full advantage of new graphics procedures and features of SAS software, including PROC SGPlot, PROC SGPANEL and ODS graphics options.

---

## We Want to Hear from You

SAS Press books are written *by SAS Users for SAS Users*. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit [sas.com/books](http://sas.com/books) to do the following:

- Sign up to review a book
- Recommend a topic
- Request authoring information
- Provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through [saspress@sas.com](mailto:saspress@sas.com) or [https://support.sas.com/author\\_feedback](https://support.sas.com/author_feedback).

SAS has many resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources: [sas.com/books](http://sas.com/books).

## About These Authors



Alex Dmitrienko, PhD, is Founder and President of Mediana Inc. He is actively involved in biostatistical research with an emphasis on multiplicity issues in clinical trials, subgroup analysis, innovative trial designs, and clinical trial optimization. Dr. Dmitrienko coauthored the first edition of *Analysis of Clinical Trials Using SAS®: A Practical Guide*, and he coedited *Pharmaceutical Statistics Using SAS®: A Practical Guide*.



Gary G. Koch, PhD, is Professor of Biostatistics and Director of the Biometrics Consulting Laboratory at the University of North Carolina at Chapel Hill. He has been active in the field of categorical data analysis for fifty years. Professor Koch teaches classes and seminars in categorical data analysis, consults in areas of statistical practice, conducts research, and trains many biostatistics students. He is coauthor of *Categorical Data Analysis Using SAS®, Third Edition*.

Learn more about these authors by visiting their author pages, where you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more:

<http://support.sas.com/dmitrienko>

<http://support.sas.com/koch>



# Chapter 1

## Model-based and Randomization-based Methods

Alex Dmitrienko (Mediana)

Gary G. Koch (University of North Carolina at Chapel Hill)

1.1	Introduction	1
1.2	Analysis of continuous endpoints	4
1.3	Analysis of categorical endpoints	20
1.4	Analysis of time-to-event endpoints	41
1.5	Qualitative interaction tests	56
1.6	References	61

This chapter discusses the analysis of clinical endpoints in the presence of influential covariates such as the trial site (center) or patient baseline characteristics. A detailed review of commonly used methods for continuous, categorical, and time-to-event endpoints, including model-based and simple randomization-based methods, is provided. The chapter describes parametric methods based on fixed and random effects models as well as nonparametric methods to perform stratified analysis of continuous endpoints. Basic randomization-based as well as exact and model-based methods for analyzing stratified categorical outcomes are presented. Stratified time-to-event endpoints are analyzed using randomization-based tests and the Cox proportional hazards model. The chapter also introduces statistical methods for assessing treatment-by-stratum interactions in clinical trials.

### 1.1 Introduction

---

Chapters 1 and 2 focus on the general statistical methods used in the analysis of clinical trial endpoints. It is broadly recognized that, when assessing the treatment effect on endpoints, it is important to perform an appropriate adjustment for important covariates such as patient baseline characteristics. The goal of an adjusted analysis is to provide an overall test of treatment effect in the presence of factors that have a significant effect on the outcome variables of interest. Two different types of factors known to influence the outcome are commonly encountered in clinical trials: *prognostic* and *non-prognostic* factors (Mehrotra, 2001). Prognostic factors are known to influence the outcome variables in a systematic way. For instance, the analysis of survival endpoints is often adjusted for prognostic factors such as patient's age and disease severity because these patient characteristics are strongly correlated with mortality. By contrast, non-prognostic factors are likely to impact the trial's outcome, but their effects do not exhibit a predictable pattern.

It is well known that treatment differences vary, sometimes dramatically, across investigational centers in multicenter clinical trials. However, the nature of center-to-center variability is different from the variability associated with patient's age or disease severity. Center-specific treatment differences are dependent on a large number of factors, e.g., geographical location, general quality of care, etc. As a consequence, individual centers influence the overall treatment difference in a fairly random manner, and it is natural to classify the center as a non-prognostic factor.

There are two important advantages of adjusted analysis over a simplistic pooled approach that ignores the influence of prognostic and non-prognostic factors. First, adjusted analyses are performed to improve the power of statistical inferences (Beach and Meier, 1989; Robinson and Jewell, 1991; Ford, and Norrie, and Ahmadi, 1995). It is well known that, by adjusting for a covariate in a linear model, one gains *precision*, which is proportional to the correlation between the covariate and outcome variable. The same is true for categorical and time-to-event endpoints (e.g., survival endpoints). Lagakos and Schoenfeld (1984) demonstrated that omitting an important covariate with a large hazard ratio dramatically reduces the efficiency of the score test in Cox proportional hazards models.

Further, failure to adjust for important covariates may introduce *bias*. Following the work of Cochran (1983), Lachin (2000, Section 4.4.3) demonstrated that the use of marginal unadjusted methods in the analysis of binary endpoints leads to biased estimates. The magnitude of the bias is proportional to the degree of treatment group imbalance within each stratum and the difference in event rates across the strata. Along the same line, Gail, Wieand, and Piantadosi (1984) and Gail, Tan, and Piantadosi (1988) showed that parameter estimates in many generalized linear and survival models become biased when relevant covariates are omitted from the regression.

## **Randomization-based and model-based methods**

---

For randomized clinical trials, there are two statistical postures for inferences concerning treatment comparisons. One is *randomization-based* with respect to the method for randomized assignment of patients to treatments, and the other is structural *model-based* with respect to assumed relationships between distributions of responses of patients and covariates for the randomly assigned treatments and baseline characteristics and measurements (Koch and Gillings, 1983). Importantly, the two postures are complementary concerning treatment comparisons, although in different ways and with different interpretations for the applicable populations. In this regard, randomization-based methods provide inferences for the randomized study population. Their relatively minimal assumptions are valid due to randomization of patients to treatments and valid observations of data before and after randomization. Model-based methods can enable inferences to a general population with possibly different distributions of baseline factors from those of the randomized study population, however, there can be uncertainty and/or controversy for the applicability of their assumptions concerning distributions of responses and their relationships to treatments and explanatory variables for baseline characteristics and measurements. This is particularly true when departures from such assumptions can undermine the validity of inferences for treatment comparisons.

For testing null hypotheses of no differences among treatments, randomization-based methods enable exact statistical tests via randomization distributions (either fully or by random sampling) without any other assumptions. In this regard, their scope includes Fisher's exact test for binary endpoints, the Wilcoxon rank sum test for ordinal endpoints, the permutation *t*-test for continuous endpoints, and the log-rank test for time-to-event end points. This class also includes the extensions

of these methods to adjust for the strata in a stratified randomization, i.e., the Mantel-Haenszel test for binary endpoints, the Van Elteren test for ordinal endpoints, and the stratified log-rank test for time-to-event endpoints. More generally, some randomization-based methods require sufficiently large sample sizes for estimators pertaining to treatment comparisons to have approximately multivariate normal distributions with essentially known covariance matrices (through consistent estimates) via central limit theory. On this basis, they provide test statistics for specified null hypotheses and/or confidence intervals. Moreover, such test statistics and confidence intervals can have randomization-based adjustment for baseline characteristics and measurements through the methods discussed in this chapter.

The class of model-based methods includes logistic regression models for settings with binary endpoints, the proportional odds model for ordinal endpoints, the multiple linear regression model for continuous endpoints, and the Cox proportional hazards model for time-to-event endpoints. Such models typically have assumptions for no interaction between treatments and the explanatory variables for baseline characteristics and measurements. Additionally, the proportional odds model has the proportional odds assumption and the proportional hazards model has the proportional hazards assumption. The multiple linear regression model relies on the assumption of homogeneous variance as well as the assumption that the model applies to the response itself or a transformation of the response, such as logarithms. Model-based methods have extensions to repeated measures data structures for multi-visit clinical trials. These methods include the repeated measures mixed model for continuous endpoints, generalized estimating equations for logistic regression models for binary and ordinal endpoints, and Poisson regression methods for time-to-event endpoints. For these extensions, the scope of assumptions pertain to the covariance structure for the responses, nature of missing data, and the extent of interactions of visits with treatments and baseline explanatory variables. See Chapter 7 for a discussion of these issues).

The main similarity of results from randomization-based and model-based methods in the analysis of clinical endpoints is the extent of statistical significance of their *p*-values for treatment comparisons. In this sense, the two classes of methods typically support similar conclusions concerning the existence of a non-null difference between treatments, with an advantage of randomization-based methods being their minimal assumptions for this purpose. However, the estimates for describing the differences between treatments from a randomization-based method pertain to the randomized population in a population-average way. But such an estimate for model-based methods homogeneously pertains to subpopulations that share the same values of baseline covariates in a subject-specific sense. For linear models, such estimates can be reasonably similar. However, for non-linear models, like the logistic regression model, proportional odds model, or proportional hazards model, they can be substantially different. Aside from this consideration, an advantage of model-based methods is that they have a straightforward structure for the assessment of homogeneity of treatment effects across patient subgroups with respect to the baseline covariates and/or measurements in the model (with respect to covariate-by-subgroup interactions). Model-based methods also provide estimates for the effects of the covariates and measurements in the model. And model-based estimates provide estimates for the effects of the covariates and measurements in the model. To summarize, the roles of randomization-based methods and model-based methods are complementary in the sense that each method is useful for the objectives that it addresses and the advantages of each method offset the limitations of the other method.

This chapter focuses on model-based and straightforward randomization-based methods commonly used in clinical trials. The methods will be applied to assess the magnitude of treatment effect on clinical endpoints in the presence of prognostic covariates. It will be assumed that covariates are nominal or ordinal and thus

can be used to define strata, which leads to a stratified analysis of relevant endpoints. Chapter 2 provides a detailed review of more advanced randomization-based methods, including the nonparametric randomization-based analysis of covariance methodology.

## Overview

Section 1.2 reviews popular ANOVA models with applications to the analysis of stratified clinical trials. Parametric stratified analyses in the continuous case are easily implemented using PROC GLM or PROC MIXED. The section also considers a popular nonparametric test for the analysis of stratified data in a non-normal setting. Linear regression models have been the focus of numerous monographs and research papers. The classical monographs of Rao (1973) and Searle (1971) provided an excellent discussion of the general theory of linear models. Milliken and Johnson (1984, Chapter 10); Goldberg and Koury (1990); and Littell, Freund, and Spector (1991, Chapter 7) discussed the analysis of stratified data in an unbalanced ANOVA setting and its implementation in SAS.

Section 1.3 reviews randomization-based (Cochran-Mantel-Haenszel and related methods) and model-based approaches to the analysis of categorical endpoints. It covers both asymptotic and exact inferences that can be implemented in PROC FREQ, PROC LOGISTIC, and PROC GENMOD. See Breslow and Day (1980); Koch and Edwards (1988); Lachin (2000); Stokes, Davis, and Koch (2000), and Agresti (2002) for a thorough overview of categorical analysis methods with clinical trial applications.

Section 1.4 discusses statistical methods used in the analysis of stratified time-to-event data. The section covers both randomization-based tests available in PROC LIFETEST and model-based tests based on the Cox proportional hazards regression implemented in PROC PHREG. Kalbfleisch and Prentice (1980); Cox and Oakes (1984); and Collett (1994) gave a detailed review of classical survival analysis methods. Allison (1995), Cantor (1997) and Lachin (2000, Chapter 9) provided an introduction to survival analysis with clinical applications and examples of SAS code.

Finally, Section 1.5 introduces popular tests for qualitative interactions. Qualitative interaction tests help understand the nature of the treatment-by-stratum interaction and identify patient populations that benefit the most from an experimental therapy. They are also often used in the context of sensitivity analyses.

The SAS code and data sets included in this chapter are available on the author's SAS Press page. See <http://support.sas.com/publishing/authors/dmitrienko.html>.

## 1.2 Analysis of continuous endpoints

---

This section reviews parametric and nonparametric analysis methods with applications to clinical trials in which the primary analysis is adjusted for important covariates, e.g., multicenter clinical trials. Within the parametric framework, we will focus on fixed and random effects models in a frequentist setting. The reader interested in alternative approaches based on conventional and empirical Bayesian methods is referred to Gould (1998).

### **EXAMPLE: Case study 1 (Multicenter depression trial)**

The following data will be used throughout this section to illustrate parametric analysis methods based on fixed and random effects models. Consider a clinical trial in patients with major depressive disorder that compares an experimental drug with a placebo. The primary efficacy measure was the change from baseline to the end of

the 9-week acute treatment phase in the 17-item Hamilton depression rating scale total score (HAMD17 score). Patient randomization was stratified by center.

A subset of the data collected in the depression trial is displayed below. Program 1.1 produces a summary of HAMD17 change scores and mean treatment differences observed at five centers.

### PROGRAM 1.1 Trial data in Case study 1

```

data hamd17;
    input center drug $ change @@;
    datalines;
100 P 18 100 P 14 100 D 23 100 D 18 100 P 10 100 P 17 100 D 18 100 D 22
100 P 13 100 P 12 100 D 28 100 D 21 100 P 11 100 P 6 100 D 11 100 D 25
100 P 7 100 P 10 100 D 29 100 P 12 100 P 12 100 P 10 100 D 18 100 D 14
101 P 18 101 P 15 101 D 12 101 D 17 101 P 17 101 P 13 101 D 14 101 D 7
101 P 18 101 P 19 101 D 11 101 D 9 101 P 12 101 D 11 102 P 18 102 P 15
102 P 12 102 P 18 102 D 20 102 D 18 102 P 14 102 P 12 102 D 23 102 D 19
102 P 11 102 P 10 102 D 22 102 D 22 102 P 19 102 P 13 102 D 18 102 D 24
102 P 13 102 P 6 102 D 18 102 D 26 102 P 11 102 P 16 102 D 16 102 D 17
102 D 7 102 D 19 102 D 23 102 D 12 103 P 16 103 P 11 103 D 11 103 D 25
103 P 8 103 P 15 103 D 28 103 D 22 103 P 16 103 P 17 103 D 23 103 D 18
103 P 11 103 P -2 103 D 15 103 D 28 103 P 19 103 P 21 103 D 17 104 D 13
104 P 12 104 P 6 104 D 19 104 D 23 104 P 11 104 P 20 104 D 21 104 D 25
104 P 9 104 P 4 104 D 25 104 D 19
;
proc sort data=hamd17;
    by drug center;
proc means data=hamd17 noprint;
    by drug center;
    var change;
    output out=summary n=n  mean=mean std=std;
data summary;
    set summary;
    format mean std 4.1;
    label drug="Drug"
    center="Center"
    n="Number of patients"
    mean="Mean HAMD17 change"
    std="Standard deviation";
proc print data=summary noobs label;
    var drug center n mean std;

```

**Output from Program 1.1**

Drug	Center	Number of patients	Mean HAMD17 change	Standard deviation
D	100	11	20.6	5.6
D	101	7	11.6	3.3
D	102	16	19.0	4.7
D	103	9	20.8	5.9
D	104	7	20.7	4.2
P	100	13	11.7	3.4
P	101	7	16.0	2.7
P	102	14	13.4	3.6
P	103	10	13.2	6.6
P	104	6	10.3	5.6

Output 1.1 lists the center-specific mean and standard deviation of the HAMD17 change scores in the two treatment groups. Note that the mean treatment differences are fairly consistent at Centers 100, 102, 103, and 104. However, Center 101 appears to be markedly different from the rest of the data.

As an aside note, it is helpful to remember that the likelihood of observing a similar treatment effect reversal by chance increases very quickly with the number of strata, and it is too early to conclude that Center 101 represents a true outlier (Senn, 1997, Chapter 14). We will discuss the problem of testing for *qualitative* treatment-by-stratum interactions in Section 1.5.

### 1.2.1 Fixed effects models

To introduce fixed effects models used in the analysis of stratified data, consider a study with a continuous endpoint that compares an experimental drug to a placebo across  $m$  strata (see Table 1.1). Suppose that the normally distributed outcome  $y_{ijk}$  observed on the  $k$ th patient in the  $j$ th stratum in the  $i$ th treatment group follows a two-way cell-means model:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}. \quad (1.1)$$

In Case study 1,  $y_{ijk}$ 's denote the reduction in the HAMD17 score in individual patients, and  $\mu_{ij}$ 's represent the mean reduction in the 10 cells defined by unique combinations of the treatment and stratum levels.

**TABLE 1.1** A two-arm clinical trial with  $m$  strata

Stratum 1			Stratum $m$		
Treatment	Number of patients	Mean	Treatment	Number of patients	Mean
Drug	$n_{11}$	$\mu_{11}$	...	Drug	$n_{1m}$
Placebo	$n_{21}$	$\mu_{21}$	...	Placebo	$n_{2m}$

The cell-means model goes back to Scheffe (1959) and has been discussed in numerous publications, including Speed, Hocking and Hackney (1978); and Milliken and Johnson (1984). Let  $n_{1j}$  and  $n_{2j}$  denote the sizes of the  $j$ th stratum in the experimental and placebo groups, respectively. Since it is uncommon to encounter empty strata in a clinical trial setting, we will assume there are no empty cells, i.e.,  $n_{ij} > 0$ . Let  $n_1$ ,  $n_2$ , and  $n$  denote the number of patients in the experimental and placebo groups and the total sample size, respectively, i.e.:

$$n_1 = \sum_{j=1}^m n_{1j}, \quad n_2 = \sum_{j=1}^m n_{2j}, \quad n = n_1 + n_2.$$

A special case of the cell-means model (1.1) is the familiar main-effects model with an interaction:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad (1.2)$$

Here,  $\mu$  denotes the overall mean; the  $\alpha$  parameters represent the treatment effects; the  $\beta$  parameters represent the stratum effects; and the  $(\alpha\beta)$  parameters are introduced to capture treatment-by-stratum variability.

Stratified data can be analyzed using several SAS procedures, including PROC ANOVA, PROC GLM, and PROC MIXED. Since PROC ANOVA supports balanced designs only, we will focus in this section on the other two procedures. PROC GLM and PROC MIXED provide the user with several analysis options for testing

the most important types of hypotheses about the treatment effect in the main-effects model (1.2). This section reviews hypotheses tested by the Type I, Type II, and Type III analysis methods. The Type IV analysis will not be discussed here because it is different from the Type III analysis only in the rare case of empty cells. The reader can find more information about Type IV analyses in Milliken and Johnson (1984) and Littell, Freund, and Spector (1991).

### Type I analysis

The Type I analysis is commonly introduced using the so-called  $R()$  notation proposed by Searle (1971, Chapter 6). Specifically, let  $R(\mu)$  denote the reduction in the error sum of squares due to fitting the mean  $\mu$ , i.e., fitting the reduced model

$$y_{ijk} = \mu + \varepsilon_{ijk}.$$

Similarly,  $R(\mu, \alpha)$  is the reduction in the error sum of squares associated with the model with the mean  $\mu$  and treatment effect  $\alpha$ , i.e.,

$$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}.$$

The difference  $R(\mu, \alpha) - R(\mu)$ , denoted by  $R(\alpha|\mu)$ , represents the additional reduction due to fitting the treatment effect after fitting the mean. It helps assess the amount of variability explained by the treatment accounting for the mean  $\mu$ . This notation is easy to extend to define other quantities such as  $R(\beta|\mu, \alpha)$ . It is important to note that  $R(\alpha|\mu)$ ,  $R(\beta|\mu, \alpha)$ , and other similar quantities are independent of restrictions imposed on parameters when they are computed from the normal equations. Therefore,  $R(\alpha|\mu)$ ,  $R(\beta|\mu, \alpha)$ , and the like are uniquely defined in any two-way classification model.

The Type I analysis is based on testing the  $\alpha$ ,  $\beta$ , and  $\alpha\beta$  factors in the main-effects model (1.2) in a sequential manner using  $R(\alpha|\mu)$ ,  $R(\beta|\mu, \alpha)$ , and  $R(\alpha\beta|\mu, \alpha, \beta)$ , respectively. Program 1.2 computes the  $F$  statistic and associated  $p$ -value for testing the difference between the experimental drug and placebo in Case study 1.

#### PROGRAM 1.2 Type I analysis of the HAMD17 changes in Case study 1

```
proc glm data=hamd17;
  class drug center;
  model change=drug|center/ss1;
  run;
```

**Output from Program 1.2**

Source	DF	Type I SS	Mean Square	F Value	Pr > F
drug	1	888.0400000	888.0400000	40.07	<.0001
center	4	87.1392433	21.7848108	0.98	0.4209
drug*center	4	507.4457539	126.8614385	5.72	0.0004

Output 1.2 lists the  $F$  statistics associated with the DRUG and CENTER effects as well as their interaction. (Recall that `drug|center` is equivalent to `drug center` `drug*center`.) Since the Type I analysis depends on the order of terms, it is important to make sure that the DRUG term is fitted first. The  $F$  statistic for the treatment comparison, represented by the DRUG term, is very large ( $F = 40.07$ ), which means that administration of the experimental drug results in a significant reduction of the HAMD17 score compared to placebo. Note that this *unadjusted analysis* ignores the effect of centers on the outcome variable.

The  $R()$  notation helps understand the structure and computational aspects of the inferences. However, as stressed by Speed and Hocking (1976), the notation might be confusing, and precise specification of the hypotheses being tested is clearly more helpful. As shown by Searle (1971, Chapter 7), the Type I  $F$  statistic for the treatment effect corresponds to the following hypothesis:

$$H_I : \frac{1}{n_1} \sum_{j=1}^m n_{1j} \mu_{1j} = \frac{1}{n_2} \sum_{j=1}^m n_{2j} \mu_{2j}.$$

It is clear that the Type I hypothesis of no treatment effect depends both on the true within-stratum means and the number of patients in each stratum.

Speed and Hocking (1980) presented an interesting characterization of the Type I, II, and III analyses that facilitates the interpretation of the underlying hypotheses. Speed and Hocking showed that the Type I analysis tests the following simple hypothesis of no treatment effect

$$H : \frac{1}{m} \sum_{j=1}^m \mu_{1j} = \frac{1}{m} \sum_{j=1}^m \mu_{2j}$$

under the condition that the  $\beta$  and  $\alpha\beta$  factors are both equal to 0. This characterization implies that the Type I analysis ignores center effects, and it is prudent to perform it when the stratum and treatment-by-stratum interaction terms are known to be negligible.

The standard ANOVA approach outlined above emphasizes hypothesis testing, and it is helpful to supplement the computed  $p$ -value for the treatment comparison with an estimate of the average treatment difference and a 95% confidence interval. The estimation procedure is closely related to the Type I hypothesis of no treatment effect. Specifically, the “average treatment difference” is estimated in the Type I framework by

$$\frac{1}{n_1} \sum_{j=1}^m n_{1j} \bar{y}_{1j} - \frac{1}{n_2} \sum_{j=1}^m n_{2j} \bar{y}_{2j}.$$

It is easy to verify from Output 1.1 and Model (1.2) that the Type I estimate of the average treatment difference in Case study 1 is equal to

$$\begin{aligned} \hat{\delta} &= \hat{\alpha}_1 - \hat{\alpha}_2 + \left( \frac{11}{50} - \frac{13}{50} \right) \hat{\beta}_1 + \left( \frac{7}{50} - \frac{7}{50} \right) \hat{\beta}_2 \\ &\quad + \left( \frac{16}{50} - \frac{14}{50} \right) \hat{\beta}_3 + \left( \frac{9}{50} - \frac{10}{50} \right) \hat{\beta}_4 + \left( \frac{7}{50} - \frac{6}{50} \right) \hat{\beta}_5 \\ &\quad + \frac{11}{50} (\widehat{\alpha\beta})_{11} + \frac{7}{50} (\widehat{\alpha\beta})_{12} + \frac{16}{50} (\widehat{\alpha\beta})_{13} + \frac{9}{50} (\widehat{\alpha\beta})_{14} + \frac{7}{50} (\widehat{\alpha\beta})_{15} \\ &\quad - \frac{13}{50} (\widehat{\alpha\beta})_{21} - \frac{7}{50} (\widehat{\alpha\beta})_{22} - \frac{14}{50} (\widehat{\alpha\beta})_{23} - \frac{10}{50} (\widehat{\alpha\beta})_{24} - \frac{6}{50} (\widehat{\alpha\beta})_{25} \\ &= \hat{\alpha}_1 - \hat{\alpha}_2 - 0.04\hat{\beta}_1 + 0\hat{\beta}_2 + 0.04\hat{\beta}_3 - 0.02\hat{\beta}_4 + 0.02\hat{\beta}_5 \\ &\quad + 0.22(\widehat{\alpha\beta})_{11} + 0.14(\widehat{\alpha\beta})_{12} + 0.32(\widehat{\alpha\beta})_{13} + 0.18(\widehat{\alpha\beta})_{14} + 0.14(\widehat{\alpha\beta})_{15} \\ &\quad - 0.26(\widehat{\alpha\beta})_{21} - 0.14(\widehat{\alpha\beta})_{22} - 0.28(\widehat{\alpha\beta})_{23} - 0.2(\widehat{\alpha\beta})_{24} - 0.12(\widehat{\alpha\beta})_{25}. \end{aligned}$$

To compute this estimate and its associated standard error, we can use the ESTIMATE statement in PROC GLM as shown in Program 1.3.

**PROGRAM 1.3 Type I estimate of the average treatment difference in Case study 1**

```
proc glm data=hamd17;
  class drug center;
  model change=drug|center/ss1;
  estimate "Trt diff"
    drug 1 -1
    center -0.04 0 0.04 -0.02 0.02
    drug*center 0.22 0.14 0.32 0.18 0.14 -0.26 -0.14 -0.28 -0.2 -0.12;
run;
```

**Output from Program 1.3**

Parameter	Estimate	Standard Error	t Value	Pr >  t
Trt diff	5.96000000	0.94148228	6.33	<.0001

Output 1.3 displays an estimate of the average treatment difference along with its standard error that can be used to construct a 95% confidence interval associated with the obtained estimate. The  $t$ -test for the equality of the treatment difference to 0 is identical to the  $F$  test for the DRUG term in Output 1.2. We can check that the  $t$  statistic in Output 1.3 is equal to the square root of the corresponding  $F$  statistic in Output 1.2. It is also easy to verify that the average treatment difference is simply the difference between the mean changes in the HAMD17 score observed in the experimental and placebo groups without any adjustment for center effects.

**Type II analysis**

In the Type II analysis, each term in the main-effects model (1.2) is adjusted for all other terms with the exception of higher-order terms that contain the term in question. Using the  $R()$  notation, the significance of the  $\alpha$ ,  $\beta$ , and  $(\alpha\beta)$  factors is tested in the Type II framework using  $R(\alpha|\mu, \beta)$ ,  $R(\beta|\mu, \alpha)$ , and  $R(\alpha\beta|\mu, \alpha, \beta)$ , respectively.

Program 1.4 computes the Type II  $F$  statistic to test the significance of the treatment effect on changes in the HAMD17 score.

**PROGRAM 1.4 Type II analysis of the HAMD17 changes in Case study 1**

```
proc glm data=hamd17;
  class drug center;
  model change=drug|center/ss2;
run;
```

**Output from Program 1.4**

Source	DF	Type II SS	Mean Square	F Value	Pr > F
drug	1	889.7756912	889.7756912	40.15	<.0001
center	4	87.1392433	21.7848108	0.98	0.4209
drug*center	4	507.4457539	126.8614385	5.72	0.0004

We see from Output 1.4 that the  $F$  statistic corresponding to the DRUG term is highly significant ( $F = 40.15$ ), which indicates that the experimental drug

significantly reduces the HAMD17 score after an adjustment for the center effect. Note that, by the definition of the Type II analysis, the presence of the interaction term in the model or the order in which the terms are included in the model do not affect the inferences with respect to the treatment effect. Thus, dropping the DRUG\*CENTER term from the model generally has little impact on the  $F$  statistic for the treatment effect. (To be precise, excluding the DRUG\*CENTER term from the model has no effect on the numerator of the  $F$  statistic but affects its denominator due to the change in the error sum of squares.)

Searle (1971, Chapter 7) demonstrated that the hypothesis of no treatment effect tested in the Type II framework has the following form:

$$H_{II} : \sum_{j=1}^m \frac{n_{1j}n_{2j}}{n_{1j} + n_{2j}} \mu_{1j} = \sum_{j=1}^m \frac{n_{1j}n_{2j}}{n_{1j} + n_{2j}} \mu_{2j}.$$

Again, as in the case of Type I analyses, the Type II hypothesis of no treatment effect depends on the number of patients in each stratum. It is interesting to note that the variance of the estimated treatment difference in the  $j$ th stratum, i.e.,  $\text{Var}(\bar{y}_{1j\cdot} - \bar{y}_{2j\cdot})$  - is inversely proportional to  $n_{1j}n_{2j}/(n_{1j} + n_{2j})$ . This means that the Type II method averages stratum-specific estimates of the treatment difference with weights proportional to the precision of the estimates.

The Type II estimate of the average treatment difference is given by

$$\left( \sum_{j=1}^m \frac{n_{1j}n_{2j}}{n_{1j} + n_{2j}} \right)^{-1} \sum_{j=1}^m \frac{n_{1j}n_{2j}}{n_{1j} + n_{2j}} (\bar{y}_{1j\cdot} - \bar{y}_{2j\cdot}). \quad (1.3)$$

For example, we can see from Output 1.1 and Model (1.2) that the Type II estimate of the average treatment difference in Case study 1 equals

$$\begin{aligned} \hat{\delta} &= \hat{\alpha}_1 - \hat{\alpha}_2 + \left( \frac{11 \times 13}{11 + 13} + \frac{7 \times 7}{7 + 7} + \frac{16 \times 14}{16 + 14} + \frac{9 \times 10}{9 + 10} + \frac{7 \times 6}{7 + 6} \right)^{-1} \times \left( \frac{11 \times 13}{11 + 13} \widehat{(\alpha\beta)}_{11} + \right. \\ &\quad \left. \frac{7 \times 7}{7 + 7} \widehat{(\alpha\beta)}_{12} + \frac{16 \times 14}{16 + 14} \widehat{(\alpha\beta)}_{13} + \frac{9 \times 10}{9 + 10} \widehat{(\alpha\beta)}_{14} + \frac{7 \times 6}{7 + 6} \widehat{(\alpha\beta)}_{15} - \frac{11 \times 13}{11 + 13} \widehat{(\alpha\beta)}_{21} - \right. \\ &\quad \left. \frac{7 \times 7}{7 + 7} \widehat{(\alpha\beta)}_{22} - \frac{16 \times 14}{16 + 14} \widehat{(\alpha\beta)}_{23} - \frac{9 \times 10}{9 + 10} \widehat{(\alpha\beta)}_{24} - \frac{7 \times 6}{7 + 6} \widehat{(\alpha\beta)}_{25} \right) \\ &= \hat{\alpha}_1 - \hat{\alpha}_2 + 0.23936 \widehat{(\alpha\beta)}_{11} + 0.14060 \widehat{(\alpha\beta)}_{12} + 0.29996 \widehat{(\alpha\beta)}_{13} + 0.19029 \widehat{(\alpha\beta)}_{14} \\ &\quad + 0.12979 \widehat{(\alpha\beta)}_{15} - 0.23936 \widehat{(\alpha\beta)}_{21} - 0.14060 \widehat{(\alpha\beta)}_{22} - 0.29996 \widehat{(\alpha\beta)}_{23} \\ &\quad - 0.19029 \widehat{(\alpha\beta)}_{24} - 0.12979 \widehat{(\alpha\beta)}_{25}. \end{aligned}$$

Program 1.5 computes the Type II estimate and its standard error using the ESTIMATE statement in PROC GLM.

### **PROGRAM 1.5    Type II estimate of the average treatment difference in Case study 1**

```
proc glm data=hamd17;
  class drug center;
  model change=drug|center/ss2;
  estimate "Trt diff"
    drug 1 -1
    drug*center 0.23936 0.14060 0.29996 0.19029 0.12979
                -0.23936 -0.14060 -0.29996 -0.19029 -0.12979;
  run;
```

Output from Program 1.5	Parameter	Estimate	Standard Error	t Value	Pr >  t
	Trt diff	5.97871695	0.94351091	6.34	<.0001

Output 1.5 shows the Type II estimate of the average treatment difference and its standard error. As in the Type I framework, the  $t$  statistic in Output 1.5 equals the square root of the corresponding  $F$  statistic in Output 1.4, which implies that the two tests are equivalent. Note also that the  $t$  statistics for the treatment comparison produced by the Type I and II analysis methods are very close in magnitude:  $t = 6.33$  in Output 1.3, and  $t = 6.34$  in Output 1.5. This similarity is not a coincidence and is explained by the fact that patient randomization was stratified by center in this trial. As a consequence,  $n_{1j}$  is close to  $n_{2j}$  for any  $j = 1, \dots, 5$ , and thus  $n_{1j}n_{2j}/(n_{1j} + n_{2j})$  is proportional to  $n_{1j}$ . The weighting schemes underlying the Type I and II tests are almost identical to each other, which causes the two methods to yield similar results. Since the Type II method becomes virtually identical to the simple Type I method when patient randomization is stratified by the covariate used in the analysis, we do not gain much from using the randomization factor as a covariate in a Type II analysis. In general, however, the standard error of the Type II estimate of the treatment difference is considerably smaller than that of the Type I estimate. Therefore, the Type II method has more power to detect a treatment effect compared to the Type I method.

As demonstrated by Speed and Hocking (1980), the Type II method tests the simple hypothesis

$$H : \frac{1}{m} \sum_{j=1}^m \mu_{1j} = \frac{1}{m} \sum_{j=1}^m \mu_{2j}$$

when the  $\alpha\beta$  factor is assumed to equal 0 (Speed and Hocking, 1980). In other words, the Type II analysis method arises naturally in trials where the treatment difference does not vary substantially from stratum to stratum.

### Type III analysis

The Type III analysis is based on a generalization of the concepts underlying the Type I and Type II analyses. Unlike these two analysis methods, the Type III methodology relies on a reparameterization of the main-effects model (1.2). The reparameterization is performed by imposing certain restrictions on the parameters in (1.2) in order to achieve a full-rank model. For example, it is common to assume that

$$\begin{aligned} \sum_{i=1}^2 \alpha_i &= 0, \quad \sum_{j=1}^m \beta_j = 0, \\ \sum_{i=1}^2 (\alpha\beta)_{ij} &= 0, \quad j = 1, \dots, m, \quad \sum_{j=1}^m (\alpha\beta)_{ij} = 0, \quad i = 1, 2. \end{aligned} \quad (1.4)$$

Once the restrictions have been imposed, one can test the  $\alpha$ ,  $\beta$ , and  $\alpha\beta$  factors using the  $R$  quantities associated with the obtained reparametrized model. (These quantities are commonly denoted by  $R^*$ .)

The introduced analysis method is more flexible than the Type I and II analyses and enables us to test hypotheses that cannot be tested using the original  $R$  quantities

(Searle, 1976; Speed and Hocking, 1976). For example, as shown by Searle (1971, Chapter 7),  $R(\alpha|\mu, \beta, \alpha\beta)$  and  $R(\beta|\mu, \alpha, \alpha\beta)$  are not meaningful when computed from the main-effects model (1.2) because they are identically equal to 0. This means that the Type I/II framework precludes us from fitting an interaction term before the main effects. By contrast,  $R^*(\alpha|\mu, \beta, \alpha\beta)$  and  $R^*(\beta|\mu, \alpha, \alpha\beta)$  associated with the full-rank reparametrized model can assume non-zero values depending on the constraints imposed on the model parameters. Thus, each term in 1.2 can be tested in the Type III framework using an adjustment for all other terms in the model.

The Type III analysis in PROC GLM and PROC MIXED assesses the significance of the  $\alpha$ ,  $\beta$ , and  $\alpha\beta$  factors using  $R^*(\alpha|\mu, \beta, \alpha\beta)$ ,  $R^*(\beta|\mu, \alpha, \alpha\beta)$  and  $R^*(\alpha\beta|\mu, \alpha, \beta)$  with the parameter restrictions given by 1.4. As an illustration, Program 1.6 tests the significance of the treatment effect on HAMD17 changes using the Type III approach.

### PROGRAM 1.6 Type III analysis of the HAMD17 changes in Case study 1

```
proc glm data=hamd17;
  class drug center;
  model change=drug|center/ss3;
  run;
```

#### Output from Program 1.6

Source	DF	Type III SS	Mean Square	F Value	Pr > F
drug	1	709.8195519	709.8195519	32.03	<.0001
center	4	91.4580063	22.8645016	1.03	0.3953
drug*center	4	507.4457539	126.8614385	5.72	0.0004

Output 1.6 indicates that the results of the Type III analysis are consistent with the Type I and II inferences for the treatment comparison. The treatment effect is highly significant after an adjustment for the center effect and treatment-by-center interaction ( $F = 32.03$ ).

The advantage of making inferences from the reparametrized full-rank model is that the Type III hypothesis of no treatment effect has the following simple form (Speed, Hocking, and Hackney, 1978):

$$H_{III} : \quad \frac{1}{m} \sum_{j=1}^m \mu_{1j} = \frac{1}{m} \sum_{j=1}^m \mu_{2j}.$$

The Type III hypothesis states that the simple average of the true stratum-specific HAMD17 change scores is identical in the two treatment groups. The corresponding Type III estimate of the average treatment difference is equal to

$$\frac{1}{m} \sum_{j=1}^m (\bar{y}_{1j} - \bar{y}_{2j}).$$

It is instructive to contrast this estimate with the Type I estimate of the average treatment difference. As was explained earlier, the idea behind the Type I approach is that individual observations are weighted equally. By contrast, the Type III method is based on weighting observations according to the size of each stratum. As a result, the Type III hypothesis involves a direct comparison of stratum means and is not affected by the number of patients in each individual stratum. To make an analogy, the Type I analysis corresponds to the U.S. House of Representatives since the number of Representatives from each state is a function of a state's population.

The Type III analysis can be thought of as a statistical equivalent of the U.S. Senate where each state sends along two Senators.

Since the Type III estimate of the average treatment difference in Case study 1 is given by

$$\widehat{\delta} = \widehat{\alpha}_1 - \widehat{\alpha}_2 + \frac{1}{5} \left[ (\widehat{\alpha\beta})_{11} + (\widehat{\alpha\beta})_{12} + (\widehat{\alpha\beta})_{13} + (\widehat{\alpha\beta})_{14} + (\widehat{\alpha\beta})_{15} \right. \\ \left. - (\widehat{\alpha\beta})_{21} - (\widehat{\alpha\beta})_{22} - (\widehat{\alpha\beta})_{23} - (\widehat{\alpha\beta})_{24} - (\widehat{\alpha\beta})_{25} \right],$$

we can compute the estimate and its standard error using the following ESTIMATE statement in PROC GLM.

### PROGRAM 1.7 Type III estimate of the average treatment difference in Case study 1

```
proc glm data=hamd17;
  class drug center;
  model change=drug|center/ss3;
  estimate "Trt diff"
    drug 1 -1
    drug*center 0.2 0.2 0.2 0.2 0.2 -0.2 -0.2 -0.2 -0.2 -0.2;
  run;
```

---

Output from Program 1.7	Parameter	Estimate	Standard	t Value	Pr >  t
			Error		
	Trt diff	5.60912865	0.99106828	5.66	<.0001

---

Output 1.7 lists the Type III estimate of the treatment difference and its standard error. Again, the significance of the treatment effect can be assessed using the  $t$  statistic shown in Output 1.7 since the associated test is equivalent to the  $F$  test for the DRUG term in Output 1.6.

### Comparison of Type I, Type II and Type III analyses

The three analysis methods introduced in this section produce identical results in any balanced data set. The situation, however, becomes much more complicated and confusing in an unbalanced setting. We need to carefully examine the available options to choose the most appropriate analysis method. The following comparison of the Type I, II, and III analyses in PROC GLM and PROC MIXED will help the reader make more educated choices in clinical trial applications.

#### Type I analysis

The Type I analysis method averages stratum-specific treatment differences with each observation receiving the same weight. Thus, the Type I approach ignores the effects of individual strata on the outcome variable. It is clear that this approach can be used only if one is not interested in adjusting for the stratum effects.

#### Type II analysis

The Type II approach amounts to comparing weighted averages of within-stratum estimates among the treatment groups. The weights are inversely proportional to the variances of stratum-specific estimates of the treatment effect. This implies that the Type II analysis is based on an optimal weighting scheme when there is no treatment-by-stratum interaction. When the treatment difference does vary

across strata, the Type II test statistic can be viewed as a weighted average of stratum-specific treatment differences with the weights equal to sample estimates of certain population parameters. For this reason, it is commonly accepted that the Type II method is the preferred way of analyzing continuous outcome variables adjusted for prognostic factors (Fleiss, 1986; Mehrotra, 2001).

Attempts to apply the Type II method to stratification schemes based on non-prognostic factors (e.g., centers) have created much controversy in the clinical trial literature. Advocates of the Type II approach maintain that centers play the same role as prognostic factors, and thus it is appropriate to carry out Type II tests in trials stratified by center as shown in Program 1.4 (Senn, 1998; Lin, 1999). Note that the outcome of the Type II analysis is unaffected by the significance of the interaction term. The interaction analysis is run separately as part of routine sensitivity analyses such as the assessment of treatment effects in various subsets and identification of outliers (Kallen, 1997; Phillips et al., 2000).

### Type III analysis

The opponents of the Type II approach argue that centers are intrinsically different from prognostic factors. Since investigative sites actively recruit patients, the number of patients enrolled at any given center is a rather arbitrary figure, and inferences driven by the sizes of individual centers are generally difficult to interpret (Fleiss, 1986). As an alternative, we can follow Yates (1934) and Cochran (1954a), who proposed to perform an analysis based on a simple average of center-specific estimates in the presence of a pronounced interaction. This unweighted analysis is equivalent to the Type III analysis of the model with an interaction term (see Program 1.6).

It is worth drawing the reader's attention to the fact that the described alternative approach based on the Type III analysis has a number of limitations:

- The Type II  $F$  statistic is generally larger than the Type III  $F$  statistic (compare Output 1.4 and Output 1.6), and thus the Type III analysis is less powerful than the Type II analysis when the treatment difference does not vary much from center to center.
- The Type III method violates the marginality principle formulated by Nelder (1977). The principle states that meaningful inferences in a two-way classification setting are to be based on the main effects  $\alpha$  and  $\beta$  adjusted for each other and on their interaction adjusted for the main effects. When we fit an interaction term before the main effects (as in the Type III analysis), the resulting test statistics depend on a totally arbitrary choice of parameter constraints. The marginality principle implies that the Type III inferences yield uninterpretable results in unbalanced cases. See Nelder (1984) and Rodriguez, Tobias, and Wolfinger (1995) for a further discussion of pros and cons of this argument.
- Weighting small and large strata equally is completely different from how we would normally perform a meta-analysis of the results observed in the strata (Senn, 2000).
- Lastly, as pointed out in several publications, sample size calculations are almost always done within the Type II framework, i.e., patients rather than centers are assumed equally weighted. As a consequence, the use of the Type III analysis invalidates the sample size calculation method. For a detailed power comparison of the weighted and unweighted approaches, see Jones et al. (1998) and Gallo (2000).

### Type III analysis with pre-testing

The described weighted and unweighted analysis methods are often combined to increase the power of the treatment comparison. As proposed by Fleiss (1986), the significance of the interaction term is assessed first, and the Type III analysis with

an interaction is performed if the preliminary test has yielded a significant outcome. Otherwise, the interaction term is removed from the model, and thus the treatment effect is analyzed using the Type II approach. The sequential testing procedure recognizes the power advantage of the weighted analysis when the treatment-by-center interaction appears to be negligible.

Most commonly, the treatment-by-center variation is evaluated using an  $F$  test based on the interaction mean square. (See the  $F$  test for the DRUG\*CENTER term in Output 1.6). This test is typically carried out at the 0.1 significance level (Fleiss, 1986). Several alternative approaches have been suggested in the literature. Bancroft (1968) proposed to test the interaction term at the 0.25 level before including it in the model. Chinchilli and Bortey (1991) described a test for consistency of treatment differences across strata based on the non-centrality parameter of an  $F$  distribution. Ciminera et al. (1993) stressed that tests based on the interaction mean square are aimed at detecting *quantitative interactions* that might be caused by a variety of factors such as measurement scale artifacts.

When applying the pre-testing strategy, we need to be aware of the fact that pre-testing leads to more frequent false-positive outcomes, which may become an issue in pivotal clinical trials. To stress this point, Jones et al. (1998) compared the described pre-testing approach with the controversial practice of pre-testing the significance of the carryover effect in crossover trials that is known to inflate the false-positive rate.

### 1.2.2 Random effects models

A popular alternative to the fixed effects modeling approach described in Section 1.2.1 is to explicitly incorporate random variation among strata in the analysis. Even though most of the discussion on center effects in the ICH guidance document entitled “Statistical principles for clinical trials” (ICH E9) treats center as a fixed effect, the guidance also encourages trialists to explore the heterogeneity of the treatment effect across centers using mixed models. The latter can be accomplished by using models with random stratum and treatment-by-stratum interaction terms. While we can argue that the selection of centers is not necessarily a random process, but treating centers as a random effect could at times help statisticians better account for between-center variability.

Random effects modeling is based on the following mixed model for the continuous outcome  $y_{ijk}$  observed on the  $k$ th patient in the  $j$ th stratum in the  $i$ th treatment group:

$$y_{ijk} = \mu + \alpha_i + b_j + g_{ij} + \varepsilon_{ijk}, \quad (1.5)$$

where  $\mu$  denotes the overall mean;  $\alpha_i$  is the fixed effect of the  $i$ th treatment;  $b_j$  and  $g_{ij}$  denote the random stratum and treatment-by-stratum interaction effects; and  $\varepsilon_{ijk}$  is a residual term. The random and residual terms are assumed to be normally distributed and independent of each other. We can see from Model (1.5) that, unlike fixed effects models, random effects models account for the variability across strata in judging the significance of the treatment effect.

Applications of mixed effects models to stratified analyses in a clinical trial context were described by several authors, including Fleiss (1986), Senn (1998), and Gallo (2000). Chakravorti and Grizzle (1975) provided a theoretical foundation for random effects modeling in stratified trials based on the familiar randomized block design framework and the work of Hartley and Rao (1967). For a detailed overview of issues related to the analysis of mixed effects models, see Searle (1992, Chapter 3). Littell, Milliken, Stroup, and Wolfinger (1996, Chapter 2) demonstrated how to use PROC MIXED in order to fit random effects models in multicenter trials.

Program 1.8 fits a random effects model to the HAMD17 data set using PROC MIXED and computes an estimate of the average treatment difference. The

DDFM=SATTERTH option in Program 1.8 requests that the degrees of freedom for the  $F$  test be computed using the Satterthwaite formula. The Satterthwaite method provides a more accurate approximation to the distribution of the  $F$  statistic in random effects models than the standard ANOVA method. It is achieved by increasing the number of degrees of freedom for the  $F$  statistic.

### **PROGRAM 1.8 Analysis of the HAMD17 changes in Case study 1 using a random effects model**

```
proc mixed data=hamd17;
  class drug center;
  model change=drug/ddfm=satterth;
  random center drug*center;
  estimate "Trt eff" drug 1 -1;
  run;
```

#### **Output from Program 1.8**

Type 3 Tests of Fixed Effects					
Effect	Num DF	Den DF	F Value	Pr > F	
drug	1	6.77	9.30	0.0194	
Estimates					
Label	Estimate	Standard Error	DF	t Value	Pr >  t
Trt eff	5.7072	1.8718	6.77	3.05	0.0194

Output 1.8 displays the  $F$  statistic ( $F = 9.30$ ) and  $p$ -value ( $p = 0.0194$ ) that are associated with the DRUG term in the random effects model as well as an estimate of the average treatment difference. The estimated treatment difference equals 5.7072 and is close to the estimates computed from fixed effects models. The standard error of the estimate (1.8718) is substantially greater than the standard error of the estimates obtained in fixed effects models (see Output 1.6). This is a penalty we have to pay for treating the stratum and interaction effects as random and reflects lack of homogeneity across the five strata in Case study 1. Note, for example, that dropping Center 101 creates more homogeneous strata and, as a consequence, reduces the standard error to 1.0442. Similarly, removing the DRUG\*CENTER term from the RANDOM statement leads to a more precise estimate of the treatment effect with the standard error of 1.0280.

In general, as shown by Senn (2000), fitting main effects as random leads to lower standard errors. However, assuming a random interaction term increases the standard error of the estimated treatment difference. Due to the lower precision of treatment effect estimates, analysis of stratified data based on models with random stratum and treatment-by-stratum effects has lower power compared to a fixed effects analysis (Gould, 1998; Jones et al., 1998).

### **1.2.3 Nonparametric tests**

This section briefly describes a nonparametric test for stratified continuous data due to van Elteren (1960). To introduce the van Elteren test, consider a clinical trial with a continuous endpoint measured in  $m$  strata. Let  $w_j$  denote the Wilcoxon

rank-sum statistic for testing the null hypothesis of no treatment effect in the  $j$ th stratum (Hollander and Wolfe, 1999, Chapter 4). Van Elteren (1960) proposed to combine stratum-specific Wilcoxon rank-sum statistics with weights inversely proportional to stratum sizes. The van Elteren statistic is given by

$$u = \sum_{j=1}^m \frac{w_j}{n_{1j} + n_{2j} + 1},$$

where  $n_{1j} + n_{2j}$  is the total number of patients in the  $j$ th stratum. To justify this weighting scheme, van Elteren demonstrated that the resulting test has asymptotically the maximum power against a broad range of alternative hypotheses. Van Elteren also studied the asymptotic properties of the testing procedure and showed that, under the null hypothesis of no treatment effect in the  $m$  strata, the test statistic is asymptotically normal.

As shown by Koch et al. (1982, Section 2.3), the van Elteren test is a member of a general family of Mantel-Haenszel mean score tests. This family also includes the Cochran-Mantel-Haenszel test for categorical outcomes discussed later in Section 1.3.1. Like other testing procedures in this family, the van Elteren test possesses an interesting and useful property: that is, its asymptotic distribution is not directly affected by the size of individual strata. As a consequence, we can rely on asymptotic  $p$ -values even in sparse stratifications as long as the total sample size is large enough. For more information about the van Elteren test and related testing procedures, see Lehmann (1975), Koch et al. (1990), and Hosmane, Shu, and Morris (1994).

#### **EXAMPLE: Case study 2 (Urinary incontinence trial)**

The van Elteren test is an alternative method of analyzing stratified continuous data when we cannot rely on standard ANOVA techniques because the underlying normality assumption is not met. As an illustration, consider a subset of the data collected in a urinary incontinence trial comparing an experimental drug to placebo over an 8-week period. The primary endpoint in the trial was a percent change from baseline to the end of the study in the number of incontinence episodes per week. Patients were allocated to three strata according to the baseline frequency of incontinence episodes<sup>1</sup>.

Program 1.9 displays a subset of the data collected in the urinary incontinence trial from Case study 2 and plots the probability distribution of the primary endpoint in the three strata.

#### **PROGRAM 1.9 Percent changes in the frequency of incontinence episodes in Case study 2**

```
data urininc;
  input therapy $ stratum @@;
  do i=1 to 10;
    input change @@;
    if (change ^= .) then output;
  end;
  drop i;
  datalines;
```

---

<sup>1</sup>This clinical trial example will be used here to illustrate a method for the analysis of non-normally distributed endpoints in the presence of a categorical stratification variable. We can think of other ways of analyzing the urinary incontinence data that might be more appropriate in this setting. For example, one can consider re-defining the primary outcome variable since a variable based on percent change from baseline makes an inefficient use of data. Further, categorizing continuous data leads to loss of power and thus the analysis described above will be inferior to an analysis which uses the baseline frequency of incontinence episodes as a continuous covariate. Yet another sensible approach is based on fitting a model that accounts for the discrete nature of incontinence episodes, e.g., a Poisson regression model for counts.

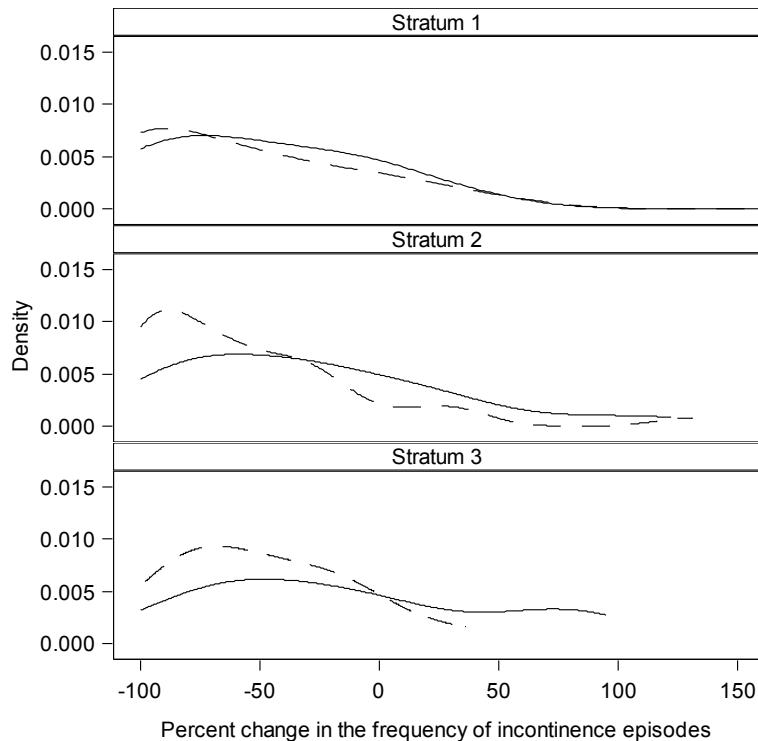
```

Placebo 1 -86 -38 43 -100 289 0 -78 38 -80 -25
Placebo 1 -100 -100 -50 25 -100 -100 -67 0 400 -100
Placebo 1 -63 -70 -83 -67 -33 0 -13 -100 0 -3
Placebo 1 -62 -29 -50 -100 0 -100 -60 -40 -44 -14
Placebo 2 -36 -77 -6 -85 29 -17 -53 18 -62 -93
Placebo 2 64 -29 100 31 -6 -100 -30 11 -52 -55
Placebo 2 -100 -82 -85 -36 -75 -8 -75 -42 122 -30
Placebo 2 22 -82 . . . . . . .
Placebo 3 12 -68 -100 95 -43 -17 -87 -66 -8 64
Placebo 3 61 -41 -73 -42 -32 12 -69 81 0 87
Drug 1 50 -100 -80 -57 -44 340 -100 -100 -25 -74
Drug 1 0 43 -100 -100 -100 -100 -63 -100 -100 -100
Drug 1 -100 -100 0 -100 -50 0 0 -83 369 -50
Drug 1 -33 -50 -33 -67 25 390 -50 0 -100 .
Drug 2 -93 -55 -73 -25 31 8 -92 -91 -89 -67
Drug 2 -25 -61 -47 -75 -94 -100 -69 -92 -100 -35
Drug 2 -100 -82 -31 -29 -100 -14 -55 31 -40 -100
Drug 2 -82 131 -60 . . . . . .
Drug 3 -17 -13 -55 -85 -68 -87 -42 36 -44 -98
Drug 3 -75 -35 7 -57 -92 -78 -69 -21 -14 .
run;

```

Figure 1.1 plots the probability distribution of the primary endpoint in the three strata. We can see from Figure 1.1 that the distribution of the primary outcome variable is consistently skewed to the right across the three strata. Since the normality assumption is clearly violated in this data set, the analysis methods described earlier in this section may perform poorly.

**Figure 1.1**  
Case study 2



The distribution of percent changes in the frequency of incontinence episodes in the experimental arm (dashed curve) and placebo arm (solid curve) by stratum.

The magnitude of treatment effect on the frequency of incontinence episodes can be assessed more reliably using a nonparametric procedure. Program 1.10 computes the van Elteren statistic to test the null hypothesis of no treatment effect in Case study 2 using PROC FREQ. The statistic is requested by including the CMH2 and SCORES=MODRIDIT options in the TABLE statement.

**PROGRAM 1.10 Analysis of percent changes in the frequency of incontinence episodes using the van Elteren test**

```
proc freq data=urininc;
  ods select cmh;
  table stratum*therapy*change/cmh2 scores=modridit;
run;
```

**Output from Program 1.10**

---

Summary Statistics for therapy by change  
Controlling for stratum

Cochran-Mantel-Haenszel Statistics (Modified Ridit Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.2505	0.0124
2	Row Mean Scores Differ	1	6.2766	0.0122

---

Output 1.10 lists two statistics produced by PROC FREQ. (Note that extraneous information has been deleted from the output using the ODS statement.) The van Elteren statistic corresponds to the row mean scores statistic labeled “Row Mean Scores Differ” and is equal to 6.2766. Since the asymptotic *p*-value is small (*p* = 0.0122), we conclude that administration of the experimental drug resulted in a significant reduction in the frequency of incontinence episodes. To compare the van Elteren test with the Type II and III analyses in the parametric ANOVA framework, Programs 1.4 and 1.6 were rerun to test the significance of the treatment effect in Case study 2. The Type II and III *F* statistics were equal to 1.4 (*p* = 0.2384) and 2.15 (*p* = 0.1446), respectively. The parametric methods were unable to detect the treatment effect in this data set due to the highly skewed distribution of the primary endpoint.

### 1.2.4 Summary

This section discussed parametric and nonparametric methods for performing stratified analyses in clinical trials with a continuous endpoint. Parametric analysis methods based on fixed and random effects models are easy to implement using PROC GLM (fixed effects only) or PROC MIXED (both fixed and random effects).

PROC GLM and PROC MIXED support three popular methods of fitting fixed effects models to stratified data known as Type I, II, and III analyses. The analysis methods are conceptually similar to each other in the sense that they are all based on averaging stratum-specific estimates of the treatment effect. The following is a quick summary of the Type I, II, and III methods:

- Each observation receives the same weight when a Type I average of stratum-specific treatment differences is computed. Therefore, the Type I approach ignores the effects of individual strata on the outcome variable.

- The Type II approach is based on a comparison of weighted averages of stratum-specific estimates of the treatment effect with the weights being inversely proportional to the variances of these estimates. The Type II weighting scheme is optimal when there is no treatment-by-stratum interaction and can also be used when treatment differences vary across strata. It is generally agreed that the Type II method is the preferred way of analyzing continuous outcome variables adjusted for prognostic factors.
- The Type III analysis method relies on a direct comparison of stratum means, which implies that individual observations are weighted according to the size of each stratum. This analysis is typically performed in the presence of a significant treatment-by-stratum interaction. It is important to remember that Type II tests are known to have more power than Type III tests when the treatment difference does not vary much from stratum to stratum.

The information about treatment differences across strata can also be combined using random effects models in which stratum and treatment-by-stratum interaction terms are treated as random variables. Random effects inferences for stratified data can be implemented using PROC MIXED. The advantage of random effects modeling is that it helps the statistician better account for between-stratum variability. However, random effects inferences are generally less powerful than inferences based on fixed effects models. This is one of the reasons why stratified analyses based on random effects models are rarely performed in a clinical trial setting.

A stratified version of the nonparametric Wilcoxon rank-sum test, known as the van Elteren test, can be used to perform inferences in a non-normal setting. It has been shown that the asymptotic distribution of the van Elteren test statistic is not directly affected by the size of individual strata, and therefore, this testing procedure performs well in the analysis of a large number of small strata.

## **1.3 Analysis of categorical endpoints**

---

This section covers analysis of categorical outcomes in clinical trials using model-based and simple randomization-based methods. It discusses both asymptotic and exact approaches, including the following:

- Randomization-based methods (Cochran-Mantel-Haenszel test)
- Minimum variance methods
- Model-based inferences

Although the examples in this section deal with the case of binary outcomes, the described analysis methods can be easily extended to a more general case of multinomial variables. SAS procedures used below automatically invoke general categorical tests when the analysis variable assumes more than two values.

Also, the section reviews methods that treat stratification factors as fixed variables. It does not cover stratified analyses based on random effects models for categorical data because they are fairly uncommon in clinical applications. For a review of tests for stratified categorical data arising within a random effects modeling framework, see Lachin (2000, Section 4.10), and Agresti and Hartzel (2000).

## Measures of association

---

There are three common measures of association used with categorical data: risk difference, relative risk, and odds ratio. To introduce these measures, consider a clinical trial designed to compare the effects of an experimental drug and placebo on the incidence of a binary event such as improvement or survival in  $m$  strata (see Table 1.2). Let  $n_{1j1}$  and  $n_{2j1}$  denote the numbers of  $j$ th stratum patients in the experimental and placebo groups, respectively, who experienced an event of interest. Similarly,  $n_{1j2}$  and  $n_{2j2}$  denote the numbers of  $j$ th stratum patients in the experimental and placebo groups, respectively, who did not experience an event of interest.

**TABLE 1.2 A two-arm clinical trial with  $m$  strata**

Stratum 1				Stratum $m$				
Treatment	Event	No event	Total	Treatment	Event	No event	Total	
Drug	$n_{111}$	$n_{112}$	$n_{11+}$	...	Drug	$n_{1m1}$	$n_{1m2}$	$n_{1m+}$
Placebo	$n_{211}$	$n_{212}$	$n_{21+}$		Placebo	$n_{2m1}$	$n_{2m2}$	$n_{2m+}$
Total	$n_{+11}$	$n_{+12}$	$n_1$		Total	$n_{+m1}$	$n_{+m2}$	$n_m$

The risk difference, relative risk, and odds ratio of observing the binary event of interest are defined as follows:

- **Risk difference.** The true event rate in  $j$ th stratum is denoted by  $\pi_{1j}$  in the experimental group and  $\pi_{2j}$  in the placebo group. Thus, the risk difference equals  $d_j = \pi_{1j} - \pi_{2j}$ . The true event rates are estimated by sample proportions  $p_{1j} = n_{1j1}/n_{1j+}$  and  $p_{2j} = n_{2j1}/n_{2j+}$ , and the risk difference is estimated by  $\hat{d}_j = p_{1j} - p_{2j}$ .
- **Relative risk.** The relative risk of observing the event in the experimental group compared to placebo group is equal to  $r_j = \pi_{1j}/\pi_{2j}$  in the  $j$ th stratum. This relative risk is estimated by  $\hat{r}_j = p_{1j}/p_{2j}$  (assuming that  $p_{2j} > 0$ ).
- **Odds ratio.** The odds of observing the event of interest in the  $j$ th stratum is  $\pi_{1j}/(1 - \pi_{1j})$  in the experimental group and  $\pi_{2j}/(1 - \pi_{2j})$  in the placebo group. The corresponding odds ratio in the  $j$ th stratum equals

$$o_j = \frac{\pi_{1j}}{1 - \pi_{1j}} / \frac{\pi_{2j}}{1 - \pi_{2j}}$$

and is estimated by

$$\hat{o}_j = \frac{p_{1j}}{1 - p_{1j}} / \frac{p_{2j}}{1 - p_{2j}}.$$

We assume here that  $p_{1j} < 1$  and  $p_{2j} > 0$ .

Since the results and their interpretation can be affected by the measure of association used in the analysis, it is important to clearly specify whether the inferences are based on risk differences, relative risks, or odds ratios.

### EXAMPLE: Case study 3 (Severe sepsis trial)

Statistical methods for the analysis of stratified clinical trials with a binary endpoint will be illustrated using the following data. A 1690-patient, placebo-controlled clinical trial was conducted to examine the effect of an experimental drug on 28-day all-cause mortality in patients with severe sepsis. Patients were assigned to one of four strata at randomization, depending on the predicted risk of mortality computed

from the APACHE II score (Knaus et al., 1985). The APACHE II score ranges from 0 to 71, and an increased score is correlated with a higher risk of death. The results observed in each of the four strata are summarized in Table 1.3<sup>2</sup>.

**TABLE 1.3 28-day mortality data in Case study 3**

Stratum	Experimental drug			Placebo		
	Dead	Alive	Total	Dead	Alive	Total
1	33	185	218	26	189	215
2	49	169	218	57	165	222
3	48	156	204	58	104	162
4	80	130	210	118	123	241

Programs 1.11 and 1.12 below summarize the survival and mortality data collected in Case study 3. Program 1.11 uses PROC FREQ to compute the risk difference, relative risk, and odds ratio of mortality in patients at a high risk of death (Stratum 4).

#### **PROGRAM 1.11 Summary of survival and mortality data in Case study 3 (Stratum 4)**

```
data sepsis;
    input stratum therapy $ outcome $ count @@;
    if outcome="Dead" then survival=0; else survival=1;
    datalines;
    1 Placebo Alive 189 1 Placebo Dead 26
    1 Drug     Alive 185 1 Drug     Dead 33
    2 Placebo Alive 165 2 Placebo Dead 57
    2 Drug     Alive 169 2 Drug     Dead 49
    3 Placebo Alive 104 3 Placebo Dead 58
    3 Drug     Alive 156 3 Drug     Dead 48
    4 Placebo Alive 123 4 Placebo Dead 118
    4 Drug     Alive 130 4 Drug     Dead 80
    ;
proc freq data=sepsis;
    where stratum=4;
    table therapy*survival/riskdiff relrisk;
    weight count;
    run;
```

#### **Output from Program 1.11**

##### **Column 1 Risk Estimates**

	Risk	ASE	(Asymptotic) 95% Confidence Limits
<hr/>			
Row 1	0.3810	0.0335	0.3153 0.4466
Row 2	0.4896	0.0322	0.4265 0.5527
Total	0.4390	0.0234	0.3932 0.4848
Difference	-0.1087	0.0465	-0.1998 -0.0176

**Difference is (Row 1 - Row 2)**

**(Exact) 95%**

<sup>2</sup>The goal of this example is to introduce statistical tests for binary outcomes stratified by a categorical variable. In general, an analysis of this type does not make the most efficient use of the data, and the trial's sponsor might need to consider alternative approaches that involve modeling the predicted risk of mortality as a continuous variable.

Confidence Limits			
Row 1	0.3150	0.4503	
Row 2	0.4249	0.5546	
Total	0.3926	0.4862	
Difference is (Row 1 - Row 2)			
Column 2 Risk Estimates			
	Risk	ASE	(Asymptotic) 95% Confidence Limits
Row 1	0.6190	0.0335	0.5534 0.6847
Row 2	0.5104	0.0322	0.4473 0.5735
Total	0.5610	0.0234	0.5152 0.6068
Difference	0.1087	0.0465	0.0176 0.1998
Difference is (Row 1 - Row 2)			
(Exact) 95% Confidence Limits			
Row 1	0.5497	0.6850	
Row 2	0.4454	0.5751	
Total	0.5138	0.6074	
Difference is (Row 1 - Row 2)			
Estimates of the Relative Risk (Row1/Row2)			
Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	0.6415	0.4404	0.9342
Cohort (Col1 Risk)	0.7780	0.6274	0.9649
Cohort (Col2 Risk)	1.2129	1.0306	1.4276

Risk statistics shown under “Column 1 Risk Estimates” in Output 1.11 represent estimated 28-day mortality rates in the experimental (Row 1) and placebo (Row 2) groups. Similarly, risk statistics under “Column 2 Risk Estimates” refer to survival rates in the two treatment groups. PROC FREQ computes both asymptotic and exact confidence intervals for the estimated rates. The estimated risk difference is  $-0.1087$ . Thus, among patients with a poor prognosis, patients treated with the experimental drug are 11% more likely to survive (in absolute terms) than those who received placebo. Note that exact confidence intervals for risk differences are quite difficult to construct (see Coe and Tamhane, 1993 for more details), and there is no exact confidence interval associated with the computed risk difference in survival or mortality rates.

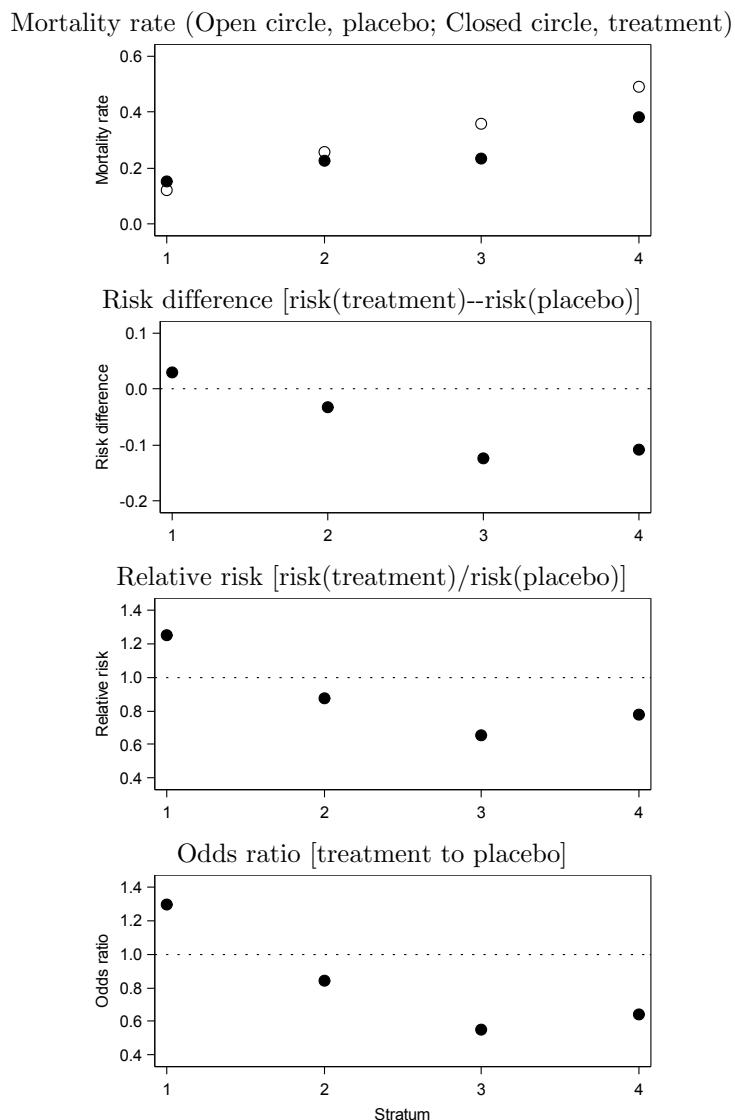
Estimates of the ratio of the odds of mortality and relative risks of survival and mortality are given under “Estimates of the Relative Risk (Row1/Row2).” The odds ratio equals 0.6415, which indicates that the odds of mortality are 36% lower in the experimental group compared to placebo in the chosen subpopulation of patients. The corresponding relative risks of survival and mortality are 1.2129 and 0.7780, respectively. The displayed 95% confidence limits are based on a normal approximation. An exact confidence interval for the odds ratio can be requested using the EXACT statement with the OR option. PROC FREQ does not currently compute exact confidence limits for relative risks.

Program 1.12 demonstrates how to use the Output Delivery System (ODS) with PROC FREQ to compute risk differences, relative risks, and odds ratios of mortality in all 4 strata.

### PROGRAM 1.12 Summary of mortality data in Case study 3 (all strata)

```
proc freq data=sepsis nopol;
  by stratum;
  table therapy*survival/riskdiff relrisk;
  ods output riskdiffcol1=riskdiff relativisks=relrisk;
  weight count;
run;
```

A summary of mortality data in Case study 3 produced by Program 1.12 is displayed in Figure 1.2. This figure shows that there was significant variability among the strata in terms of 28-day mortality rates. The absolute reduction in



**Figure 1.2**  
Case study 3

*A summary of mortality data in Case study 3 accompanied by three measures of treatment effect.*

mortality in the experimental arm compared to placebo varied from  $-3.0\%$  in Stratum 1 to  $12.3\%$  in Stratum 3. The treatment effect was most pronounced in patients with a poor prognosis at study entry, i.e., patients in Stratum 3 and 4.

### 1.3.1 Asymptotic randomization-based tests

Fleiss (1981, Chapter 10) described a general method for performing stratified analyses that goes back to Cochran (1954a). Fleiss applied it to the case of binary outcomes. Let  $a_j$  and  $s_j^2$  denote the estimates of a certain measure of association between the treatment and binary outcome and its sample variance in the  $j$ th stratum, respectively. Assume that the measure of association is chosen in such a way that it equals 0 when the treatment difference is 0. Also,  $w_j$  will denote the reciprocal of the sample variance, i.e.,  $w_j = 1/s_j^2$ . The total chi-square statistic

$$\chi_T^2 = \sum_{j=1}^m w_j a_j^2$$

can be partitioned into a chi-square statistic  $\chi_H^2$  for testing the degree of homogeneity among the strata and a chi-square statistic  $\chi_A^2$  for testing the significance of overall association across the strata given by

$$\chi_H^2 = \sum_{j=1}^m w_j (a_j - \hat{a})^2, \quad (1.6)$$

$$\chi_A^2 = \left( \sum_{j=1}^m w_j \right)^{-1} \left( \sum_{j=1}^m w_j a_j \right)^2, \quad (1.7)$$

where

$$\hat{a} = \left( \sum_{j=1}^m w_j \right)^{-1} \sum_{j=1}^m w_j a_j \quad (1.8)$$

is the associated minimum variance estimate of the degree of association averaged across the  $m$  strata. Under the null hypothesis of homogeneous association,  $\chi_H^2$  asymptotically follows a chi-square distribution with  $m - 1$  degrees of freedom. Similarly, under the null hypothesis that the average association between the treatment and binary outcome is zero,  $\chi_A^2$  is asymptotically distributed as chi-square with 1 degree of freedom.

The described method for testing hypotheses of homogeneity and association in a stratified setting can be used to construct a large number of useful tests. For example, if  $a_j$  is equal to a standardized treatment difference in the  $j$ th stratum,

$$a_j = \frac{\hat{d}_j}{\bar{p}_j(1-\bar{p}_j)}, \text{ where } \bar{p}_j = n_{+j1}/n_j \text{ and } \hat{d}_j = p_{1j} - p_{2j}, \quad (1.9)$$

then

$$w_j = \bar{p}_j(1-\bar{p}_j) \frac{n_{1j} + n_{2j}}{n_{1j} + n_{2j}}$$

and the associated chi-square test of overall association based on  $\chi_A^2$  is equivalent to a test for stratified binary data proposed by Cochran (1954b) and is asymptotically equivalent to a test developed by Mantel and Haenszel (1959). Due to their similarity, it is common to collectively refer to the two tests as the Cochran-Mantel-Haenszel (CMH) procedure. Since  $a_j$  in (1.9) involves the estimated risk difference  $\hat{d}_j$ , the

CMH procedure tests the degree of association with respect to the risk differences  $d_1, \dots, d_m$  in the  $m$  strata. The estimate of the average risk difference corresponding to the CMH test is given by

$$\hat{d} = \left( \sum_{j=1}^m \bar{p}_j (1 - \bar{p}_j) \frac{n_{1j+} + n_{2j+}}{n_{1j+} + n_{2j+}} \right)^{-1} \sum_{j=1}^m \frac{n_{1j+} + n_{2j+}}{n_{1j+} + n_{2j+}} \hat{d}_j. \quad (1.10)$$

It is interesting to compare this estimate to the Type II estimate of the average treatment effect in the continuous case (see Section 1.2.1). The stratum-specific treatment differences  $\hat{d}_1, \dots, \hat{d}_m$ , are averaged in the CMH estimate with the same weights as in the Type II estimate. Thus, we can think of the CMH procedure as an extension of the Type II testing method to trials with a binary outcome. Although unweighted estimates corresponding to the Type III method have been mentioned in the literature, they are rarely used in the analysis of stratified trials with a categorical outcome and are not implemented in SAS.

We can use the general method described by Fleiss (1981, Chapter 10) to construct estimates and associated tests for overall treatment effect based on relative risks and odds ratios. Relative risks and odds ratios need to be transformed before the method is applied because they are equal to 1 in the absence of treatment effect. Most commonly, a log transformation is used to ensure that  $a_j = 0$ ,  $j = 1, \dots, m$ , when the stratum-specific treatment differences are equal to 0.

The minimum variance estimates of the average log relative risk and log odds ratio are based on the formula (1.8) with

$$a_j = \log \hat{r}_j, \quad w_j = \left[ \left( \frac{1}{n_{1j1}} - \frac{1}{n_{1j+}} \right) + \left( \frac{1}{n_{2j1}} - \frac{1}{n_{2j+}} \right) \right]^{-1} \text{ (log relative risk)}, \quad (1.11)$$

$$a_j = \log \hat{o}_j, \quad w_j = \left( \frac{1}{n_{1j1}} + \frac{1}{n_{1j2}} + \frac{1}{n_{2j1}} + \frac{1}{n_{2j2}} \right)^{-1} \text{ (log odds ratio)}. \quad (1.12)$$

The corresponding estimates of the average relative risk and odds ratio are computed using exponentiation. Adopting the PROC FREQ terminology, we will refer to these estimates as *logit-adjusted* estimates and denote them by  $\hat{r}_L$  and  $\hat{o}_L$ .

It is instructive to compare the logit-adjusted estimates  $\hat{r}_L$  and  $\hat{o}_L$  with estimates of the average relative risk and odds ratio proposed by Mantel and Haenszel (1959). The Mantel-Haenszel estimates, denoted by  $\hat{r}_{MH}$  and  $\hat{o}_{MH}$ , can also be expressed as weighted averages of stratum-specific relative risks and odds ratios:

$$\hat{r}_{MH} = \left( \sum_{j=1}^m w_j \right)^{-1} \sum_{j=1}^m w_j \hat{r}_j, \text{ where } w_j = \frac{n_{2j1} n_{1j+}}{n_j},$$

$$\hat{o}_{MH} = \left( \sum_{j=1}^m w_j \right)^{-1} \sum_{j=1}^m w_j \hat{o}_j, \text{ where } w_j = \frac{n_{2j1} n_{1j2}}{n_j}.$$

Note that weights in  $\hat{r}_{MH}$  and  $\hat{o}_{MH}$  are not inversely proportional to sample variances of the stratum-specific estimates. Thus,  $\hat{r}_{MH}$  and  $\hat{o}_{MH}$  do not represent minimum variance estimates. Despite this property, the Mantel-Haenszel estimates are generally comparable to the logit-adjusted estimates  $\hat{r}_L$  and  $\hat{o}_L$  in terms of precision. Also, as shown by Breslow (1981),  $\hat{r}_{MH}$  and  $\hat{o}_{MH}$  are attractive in applications because their mean square error is always less than that of the logit-adjusted estimates  $\hat{r}_L$  and  $\hat{o}_L$ . Further, Breslow (1981) and Greenland and Robins (1985)

studied the asymptotic behavior of the Mantel-Haenszel estimates and demonstrated that, unlike the logit-adjusted estimates, they perform well in sparse stratifications.

The introduced estimates of the average risk difference, relative risk, and odds ratio, as well as associated test statistics, are easy to obtain using PROC FREQ. Program 1.13 carries out the CMH test in Case study 3 controlling for the baseline risk of mortality represented by the STRATUM variable. The program also computes the logit-adjusted and Mantel-Haenszel estimates of the average relative risk and odds ratio. Note that the order of the variables in the TABLE statement is very important: the stratification factor is followed by the other two variables.

### PROGRAM 1.13 Average association between treatment and survival in Case study 3

```
proc freq data=sepsis;
  table stratum*therapy*survival/cmh;
  weight count;
  run;
```

#### Output from Program 1.13

---

Summary Statistics for therapy by outcome  
Controlling for stratum

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	6.9677	0.0083
2	Row Mean Scores Differ	1	6.9677	0.0083
3	General Association	1	6.9677	0.0083

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value
---------------	--------	-------

Case-Control (Odds Ratio)	Mantel-Haenszel	0.7438
	Logit	0.7426

Cohort (Col1 Risk)	Mantel-Haenszel	0.8173
	Logit	0.8049

Cohort (Col2 Risk)	Mantel-Haenszel	1.0804
	Logit	1.0397

Type of Study	Method	95% Confidence Limits	
---------------	--------	-----------------------	--

Case-Control (Odds Ratio)	Mantel-Haenszel	0.5968	0.9272
	Logit	0.5950	0.9267

Cohort (Col1 Risk)	Mantel-Haenszel	0.7030	0.9501
	Logit	0.6930	0.9349

Cohort (Col2 Risk)	Mantel-Haenszel	1.0198	1.1447
	Logit	0.9863	1.0961

Breslow-Day Test for  
Homogeneity of the Odds Ratios

Chi-Square	6.4950
DF	3
Pr > ChiSq	0.0899

Output 1.13 shows that the CMH statistic for association between treatment and survival adjusted for the baseline risk of death equals 6.9677 and is highly significant ( $p = 0.0083$ ). This means that there is a significant overall increase in survival across the 4 strata in patients treated with the experimental drug.

The central panel of Output 1.13 lists the Mantel-Haenszel and logit-adjusted estimates of the average relative risk and odds ratio as well as the associated asymptotic 95% confidence intervals. The estimates of the odds ratio of mortality, shown under “Case-Control (Odds Ratio,)” are 0.7438 (Mantel-Haenszel estimate  $\hat{o}_{MH}$ ) and 0.7426 (logit-adjusted estimate  $\hat{o}_L$ ). The estimates indicate that the odds of mortality adjusted for the baseline risk of mortality are about 26% lower in the experimental group compared to placebo. The estimates of the average relative risk of mortality, given under “Cohort (Col1 Risk,)” are 0.8173 (Mantel-Haenszel estimate  $\hat{r}_{MH}$ ) and 0.8049 (logit-adjusted estimate  $\hat{r}_L$ ). Since the Mantel-Haenszel estimate is known to minimize the mean square error, it is generally more reliable than the logit-adjusted estimate. Using the Mantel-Haenszel estimate, the experimental drug reduces the 28-day mortality rate by 18% (in relative terms) compared to placebo. The figures shown under “Cohort (Col2 Risk)” are the Mantel-Haenszel and logit-adjusted estimates of the average relative risk of survival.

As was mentioned above, confidence intervals for the Mantel-Haenszel estimates  $\hat{r}_{MH}$  and  $\hat{o}_{MH}$  are comparable to those based on the logit-adjusted estimates  $\hat{r}_L$  and  $\hat{o}_L$ . The 95% confidence interval associated with  $\hat{o}_L$  is given by (0.5950, 0.9267) and is slightly wider than the 95% confidence interval associated with  $\hat{o}_{MH}$  given by (0.5968, 0.9272). However, the 95% confidence interval associated with  $\hat{r}_L$  (0.6930, 0.9349) is tighter than the 95% confidence interval for  $\hat{r}_{MH}$  (0.7030, 0.9501). Note that the confidence intervals are computed from a very large sample. So it is not surprising that the difference between the two estimation methods is very small.

Finally, the bottom panel of Output 1.13 displays the Breslow-Day chi-square statistic that can be used to examine whether the odds ratio of mortality is homogeneous across the 4 strata. See Breslow and Day (1980, Section 4.4) for details. The Breslow-Day  $p$ -value equals 0.0899 and suggests that the stratum-to-stratum variability in terms of the odds ratio is not very large. It is sometimes stated in the clinical trial literature that the CMH statistic needs to be used with caution when the Breslow-Day test detects significant differences in stratum-specific odds ratios. As pointed out by Agresti (2002, Section 6.3), the CMH procedure is valid and produces a meaningful result even if odds ratios differ from strata to strata as long as no pronounced qualitative interaction is present.

It is important to remember that the Breslow-Day test is specifically formulated to compare stratum-specific odds ratios. The homogeneity of relative differences or relative risks can be assessed using other testing procedures - for example, homogeneity tests based on the framework described by Fleiss (1981, Chapter 10) or the simple interaction test proposed by Mehrotra (2001). We can also make use of tests for qualitative interaction proposed by Gail and Simon (1985) and Ciminera et al. (1993). See Section 1.5. Note that the tests for qualitative interaction are generally much more conservative than the Breslow-Day test.

### 1.3.2 Exact randomization-based tests

It is common in the categorical analysis literature to examine the asymptotic behavior of stratified tests under the following two scenarios:

- **Large-strata asymptotics.** The total sample size  $n$  is assumed to increase while the number of strata  $m$  remains fixed.
- **Sparse-data asymptotics.** The total sample size is assumed to grow with the number of strata.

The majority of estimation and hypothesis testing procedures used in the analysis of stratified categorical data perform well only in a large-strata asymptotic setting. For example, the logit-adjusted estimates of the average relative risk and odds ratio as well as the Breslow-Day test require that all strata contain a sufficiently large number of data points.

It was shown by Birch (1964) that the asymptotic theory for the CMH test is valid under sparse-data asymptotics. In other words, the CMH statistic follows a chi-square distribution even in the presence of a large number of small strata. Mantel and Fleiss (1980) studied the accuracy of the chi-square approximation and devised a simple rule to confirm the adequacy of this approximation in practice. It is appropriate to compute the CMH  $p$ -value from a chi-square distribution with 1 degree of freedom if both

$$\sum_{i=1}^m \left( \frac{n_{1i} + n_{+i1}}{n_i} - \max(0, n_{+i1} - n_{2i+}) \right) \text{ and } \sum_{i=1}^m \left( \min(n_{1i+}, n_{+i1}) - \frac{n_{1i} + n_{+i1}}{n_i} \right)$$

exceed 5. See Breslow and Day (1980, Section 4.4) or Koch and Edwards (1988) for more details.

The Mantel-Fleiss criterion is met with a wide margin in all reasonably large studies and needs to be checked only when most of the strata have low patient counts. As an illustration, consider a subset of the trial database in Case study 3, which includes the data collected at three centers (see Table 1.4).

**TABLE 1.4** 28-day mortality data from Case study 3 at three selected centers

Center	Experimental drug			Placebo		
	Alive	Dead	Total	Alive	Dead	Total
1	4	0	4	2	2	4
2	3	1	4	1	2	3
3	3	0	3	3	2	5

It is easy to verify that the Mantel-Fleiss criterion is not met for this subset because

$$\sum_{i=1}^3 \left( \frac{n_{1i} + n_{+i1}}{n_i} - \max(0, n_{+i1} - n_{2i+}) \right) = 3.464,$$

$$\sum_{i=1}^3 \left( \min(n_{1i+}, n_{+i1}) - \frac{n_{1i} + n_{+i1}}{n_i} \right) = 3.536.$$

When the Mantel-Fleiss criterion is not satisfied, we can resort to exact stratified tests. Although PROC FREQ supports exact inferences in simple binary settings, it does not currently implement exact tests or compute exact confidence intervals for stratified binary data described in the literature (Agresti, 2001). As shown by Westfall et al. (1999, Chapter 12), exact inferences for binary outcomes can be performed by carrying out the Cochran-Armitage permutation test available in PROC MULTTEST. The Cochran-Armitage test is ordinarily used for assessing the strength of a linear relationship between a binary response variable and a continuous covariate. It is known that the Cochran-Armitage permutation test simplifies to the Fisher exact test in the case of two treatment groups. Thus, we can use a stratified version of the Cochran-Armitage permutation test in PROC MULTTEST to carry out the exact Fisher test for average association between treatment and survival in Table 1.4.

Program 1.14 carries out the CMH test using PROC FREQ and also computes an exact  $p$ -value from the Cochran-Armitage permutation test using PROC MULTTEST. The Cochran-Armitage test is requested by the CA option in the

TEST statement of PROC MULTTEST. The PERMUTATION option in the TEST statement tells PROC MULTTEST to perform enumeration of all permutations using the multivariate hypergeometric distribution in small strata (stratum size is less than or equal to the specified PERMUTATION parameter) and use a continuity-corrected normal approximation otherwise.

**PROGRAM 1.14 Average association between treatment and survival at the three selected centers in Case study 3**

```
data sepsis1;
    input center therapy $ outcome $ count @@;
    if outcome="Dead" then survival=0; else survival=1;
    datalines;
    1 Placebo Alive 2 1 Placebo Dead 2
    1 Drug     Alive 4 1 Drug     Dead 0
    2 Placebo Alive 1 2 Placebo Dead 2
    2 Drug     Alive 3 2 Drug     Dead 1
    3 Placebo Alive 3 3 Placebo Dead 2
    3 Drug     Alive 3 3 Drug     Dead 0
    ;
proc freq data=sepsis1;
    table center*therapy*survival/cmh;
    weight count;
proc multtest data=sepsis1;
    class therapy;
    freq count;
    strata center;
    test ca(survival/permuation=20);
    run;
```

**Output from Program 1.14**

```
The FREQ Procedure
Summary Statistics for therapy by outcome
Controlling for center

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic      Alternative Hypothesis      DF      Value      Prob
-----
1              Nonzero Correlation        1      4.6000      0.0320
2              Row Mean Scores Differ   1      4.6000      0.0320
3              General Association       1      4.6000      0.0320

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study      Method      Value
-----
Case-Control      Mantel-Haenszel      0.0548
(Odds Ratio)      Logit **          0.1552

Cohort            Mantel-Haenszel      0.1481
(Col1 Risk)       Logit **          0.3059

Cohort            Mantel-Haenszel      1.9139
(Col2 Risk)       Logit             1.8200

Type of Study      Method      95% Confidence Limits
-----
```

Case-Control (Odds Ratio)	Mantel-Haenszel Logit **	0.0028 0.0223	1.0801 1.0803
Cohort (Col1 Risk)	Mantel-Haenszel Logit **	0.0182 0.0790	1.2048 1.1848
Cohort (Col2 Risk)	Mantel-Haenszel Logit	1.0436 1.0531	3.5099 3.1454

\*\* These logit estimators use a correction of 0.5 in every cell of those tables that contain a zero.

#### The Multtest Procedure

##### Model Information

Test for discrete variables:	Cochran-Armitage
Exact permutation distribution used:	Everywhere
Tails for discrete tests:	Two-tailed
Strata weights:	Sample size

##### p-Values

Variable	Contrast	Raw
surv	Trend	0.0721

Output 1.14 displays the CMH *p*-value as well as the Cochran-Armitage *p*-value for association between treatment and survival in the three selected centers. Since the PERMUTATION parameter specified in PROC MULTTEST is greater than all three stratum totals, the computed Cochran-Armitage *p*-value is exact.

It is important to contrast the *p*-values produced by the CMH and Cochran-Armitage permutation tests. The CMH *p*-value equals 0.0320 and is thus significant at the 5% level. Since the Mantel-Fleiss criterion is not satisfied due to very small cell counts, the validity of the CMH test is questionable. It is prudent to examine the *p*-value associated with the exact Cochran-Armitage test. The exact *p*-value (0.0721) is more than twice as large as the CMH *p*-value and indicates that the adjusted association between treatment and survival is unlikely to be significant.

Since PROC MULTTEST efficiently handles permutation-based inferences in large data sets, the described exact test for stratified binary outcomes can be easily carried out in data sets with thousands of observations. As an illustration, Program 1.15 computes the exact Cochran-Armitage *p*-value for average association between treatment and survival in Case study 3.

#### PROGRAM 1.15 Exact test for average association between treatment and survival in Case study 3

```
proc multtest data=sepsis;
  class therapy;
  freq count;
  strata stratum;
  test ca(survival/permuation=500);
  run;
```

Output from Program 1.15		p-Values
Variable	Contrast	Raw
surv	Trend	0.0097

It is easy to see from Table 1.3 that the PERMUTATION parameter used in Program 1.15 is greater than the size of each individual stratum in Case study 3. This means that PROC MULTTEST enumerated all possible permutations in the four strata and that the Cochran-Armitage  $p$ -value shown in Output 1.15 is exact. Note that the exact  $p$ -value equals 0.0097 and is close to the asymptotic CMH  $p$ -value from Output 1.13 ( $p = 0.0083$ ). One additional advantage of using the exact Cochran-Armitage test in PROC MULTTEST is that a one-sided  $p$ -value can be easily requested by adding the LOWERTAILED option after PERMUTATION=500.

### 1.3.3 Minimum risk tests

Optimal properties of the CMH test have been extensively studied in the literature. Radhakrishna (1965) provided a detailed analysis of stratified tests and demonstrated that the weighting strategy used in the CMH procedure works best (in terms of the power to detect a treatment difference) when odds ratios of an event of interest are constant across strata. This weighting strategy (known as the SSIZE strategy) may not be very effective when this assumption is not met. This happens, for example, when a constant multiplicative or constant additive treatment effect is observed (in other words, when strata are homogeneous with respect to the relative risk or risk difference). However, as demonstrated by Radhakrishna (1965), we can easily set up an asymptotically optimal test under these alternative assumptions by using a different set of stratum-specific weights (see also Lachin, 2000, Section 4.7). For example, an optimal test for the case of a constant risk difference (known as the INVAR test) is based on weights that are inversely proportional to the variances of stratum-specific estimates of treatment effect (expressed in terms of risk difference).

Despite the availability of these optimal tests, we can rarely be certain that the pre-specified test is the most efficient one since it is impossible to tell if the treatment difference is constant on a multiplicative, additive, or any other scale until the data have been collected. In order to alleviate the described problem, several authors discussed ways to minimize the power loss that can occur under the worst possible configuration of stratum-specific parameters. Gastwirth (1985) demonstrated how to construct *maximin efficiency robust tests* that maximize the minimum efficiency in a broad class of stratified testing procedures. Mehrotra and Railkar (2000) introduced a family of *minimum risk* tests that minimize the mean square error of the associated estimate of the overall treatment difference. The minimum risk procedures rely on data-driven stratum-specific weights  $w_1, \dots, w_m$  given by

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} \beta_1 + \alpha_1 \hat{d}_1 & \alpha_1 \hat{d}_2 & \dots & \alpha_1 \hat{d}_m \\ \alpha_2 \hat{d}_1 & \beta_2 + \alpha_2 \hat{d}_2 & \dots & \alpha_2 \hat{d}_m \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_m \hat{d}_1 & \alpha_m \hat{d}_2 & \dots & \beta_m + \alpha_m \hat{d}_m \end{bmatrix}^{-1} \begin{bmatrix} 1 + \alpha_1 \gamma/n \\ 1 + \alpha_2 \gamma/n \\ \vdots \\ 1 + \alpha_m \gamma/n \end{bmatrix},$$

where

$$\begin{aligned}\hat{d}_j &= p_{1j} - p_{2j}, \quad \alpha_i = \hat{d}_i \sum_{j=1}^m V_j^{-1} - \sum_{j=1}^m \hat{d}_j V_j^{-1}, \quad \gamma = \sum_{j=1}^m n_j \hat{d}_j, \\ \beta_i &= V_i \sum_{j=1}^m V_j^{-1}, \quad V_i = \frac{p_{1j}(1-p_{1j})}{n_{1j+}} + \frac{p_{2j}(1-p_{2j})}{n_{2j+}}.\end{aligned}$$

Once the weights have been calculated, the minimum risk estimate of the average treatment difference is computed,

$$\hat{d}_{MR} = \sum_{j=1}^m w_j \hat{d}_j,$$

and the minimum risk test of association across the  $m$  strata is conducted based on the following test statistic:

$$z_{MR} = \left( \sum_{j=1}^m w_j^2 V_j^* \right)^{-1/2} \left[ |\hat{d}_{MR}| - \frac{3}{16} \left( \sum_{j=1}^m \frac{n_{1j+} n_{2j+}}{n_{1j+} + n_{2j+}} \right)^{-1} \right],$$

where

$$V_j^* = \bar{p}_j(1-\bar{p}_j) \frac{n_{1j+} n_{2j+}}{n_{1j+} + n_{2j+}}$$

is the sample variance of the estimated treatment difference in the  $j$ th stratum under the null hypothesis. Assuming the null hypothesis of no treatment difference, the test statistic  $z_{MR}$  is asymptotically normally distributed. Mehrotra and Railkar (2000) showed via simulations that the normal approximation can be used even when the stratum sample sizes are fairly small, i.e., when  $n_j \geq 10$ .

The principal advantage of the minimum risk test for the strength of average association in a stratified binary setting is that it is more robust than the optimal tests constructed under the assumption of homogeneous odds ratios or risk differences. Unlike the INVAR and SSIZE procedures that are quite vulnerable to deviations from certain optimal configurations of stratum-specific treatment differences, the minimum risk procedure displays much less sensitivity to those configurations. As pointed out by Mehrotra and Railkar (2000), this is a “minimum regret” procedure that minimizes the potential power loss that can occur in the worst-case scenario. To illustrate this fact, Mehrotra and Railkar showed that the minimum risk test is more powerful than the SSIZE test when the latter is not the most efficient test, e.g., when the risk differences (rather than the odds ratios) are constant from stratum to stratum. Likewise, the minimum risk test demonstrates a power advantage over the INVAR test that is derived under the assumption of homogeneous risk differences when this assumption is not satisfied. This means that the minimum risk strategy serves as a viable alternative to the optimal tests identified by Radhakrishna (1965) when there is little *a priori* information on how the treatment difference varies across the strata.

Program 1.16 uses the minimum risk strategy to test for association between treatment and survival in Case study 3. The program computes the minimum risk estimate of the average treatment difference and carries out the minimum risk test for association (as well as the INVAR and SSIZE tests) by invoking the %MinRisk macro. The %MinRisk macro assumes that the input data set includes variables named EVENT1 (number of events of interest in Treatment group 1), EVENT2 (number of events of interest in Treatment group 2) and similarly defined NOEVENT1 and NOEVENT2 with one record per stratum. The EVENT1 and NOEVENT1 variables

in the SEPSIS2 data set below capture the number of survivors and non-survivors in the experimental group. Likewise, the EVENT2 and NOEVENT2 variables contain the number of survivors and non-survivors in the placebo group.

**PROGRAM 1.16 Minimum risk test for association between treatment and survival in Case study 3**

```
data sepsis2;
    input event1 noevent1 event2 noevent2 @@;
    datalines;
    185 33 189 26
    169 49 165 57
    156 48 104 58
    130 80 123 118
    ;
%MinRisk(dataset=sepsis2);
```

---

<b>Output from Program 1.16</b>	MINRISK Estimate   Statistic   P-value  0.0545      2.5838    0.0098
	INVAR Estimate   Statistic   P-value  0.0391      1.9237    0.0544
	SSIZE Estimate   Statistic   P-value  0.0559      2.6428    0.0082

---

Output 1.16 lists the estimates of the average difference in survival between the experimental drug and placebo and associated *p*-values produced by the minimum risk, INVAR, and SSIZE procedures. The estimate of the average treatment difference produced by the minimum risk method (0.0545) is very close in magnitude to the SSIZE estimate (0.0559). As a consequence, the minimum risk and SSIZE test statistics and *p*-values are also very close to each other. Note that, since the SSIZE testing procedure is asymptotically equivalent to the CMH procedure, the *p*-value generated by the SSIZE method is virtually equal to the CMH *p*-value shown in Output 1.13 (*p* = 0.0083).

The INVAR estimate of the overall difference is biased downward, and the associated test of the hypothesis of no difference in survival yields a *p*-value that is greater than 0.05. The INVAR testing procedure is less powerful than the SSIZE procedure in this example because the odds ratios are generally more consistent across the strata than the risk differences in survival.

Although the minimum risk test is slightly less efficient than the SSIZE test in this scenario, it is important to keep in mind that the minimum risk approach is more robust than the other two approaches in the sense that it is less dependent on the pattern of treatment effects across strata.

### 1.3.4 Asymptotic model-based tests

Model-based estimates and tests present an alternative to the randomization-based procedures introduced in the first part of this section. Model-based methods are closely related to the randomization-based procedures and address the same problem of testing for association between treatment and outcome controlling for important covariates. The difference is that this testing problem is now embedded in a modeling framework. The outcome variable is modeled as a function of selected covariates (treatment effect as well as various prognostic and non-prognostic factors, and an inferential method is applied to estimate model parameters and test associated hypotheses. One of the advantages of model-based methods is that we can compute adjusted estimates of the treatment effect in the presence of continuous covariates whereas randomization-based methods require a contingency table setup, i.e., they can be used only with categorical covariates.

The current section describes asymptotic maximum likelihood inferences based on a logistic regression model, while Section 1.3.5 discusses exact permutation-based inferences. As before, we will concentrate on the case of binary outcome variables. Refer to Stokes, Davis, and Koch (2000, Chapters 8 and 9), Lachin (2000, Chapter 7) and Agresti (2002, Chapters 5 and 7), for a detailed overview of maximum likelihood methods in logistic regression models with SAS examples.

Model-based inferences in stratified categorical data can be implemented using several SAS procedures, including PROC LOGISTIC, PROC GENMOD, PROC PROBIT, PROC CATMOD, and PROC NLMIXED. Some of these procedures are more general and provide the user with a variety of statistical modeling tools. For example, PROC GENMOD was introduced to support normal, binomial, Poisson, and other generalized linear models. And PROC NLMIXED enables the user to fit a large number of nonlinear mixed models. The others, e.g., PROC LOGISTIC and PROC PROBIT, deal with a rather narrow class of models. However, as more specialized procedures often do, they support more useful features. This section will focus mainly on one of these procedures that is widely used to analyze binary data (PROC LOGISTIC) and will briefly describe some of the interesting features of another popular procedure (PROC GENMOD).

Program 1.17 uses PROC LOGISTIC to analyze average association between treatment and survival, controlling for the baseline risk of death in Case study 3.

#### PROGRAM 1.17

#### Maximum likelihood analysis of average association between treatment and survival in Case study 3 using PROC LOGISTIC

```
proc logistic data=sepsis;
  class therapy stratum;
  model survival=therapy stratum/clodds=pl;
  freq count;
  run;
```

---

#### Output from Program 1.17

##### Type III Analysis of Effects

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
therapy	1	6.9635	0.0083
stratum	3	97.1282	<.0001

Analysis of Maximum Likelihood Estimates

Parameter			Standard	Wald	
	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.0375	0.0585	314.8468	<.0001
therapy Drug	1	-0.1489	0.0564	6.9635	0.0083
stratum 1	1	-0.8162	0.1151	50.2692	<.0001
stratum 2	1	-0.1173	0.0982	1.4259	0.2324
stratum 3	1	0.1528	0.1006	2.3064	0.1288

Odds Ratio Estimates					
Effect			Point	95% Wald	
			Estimate	Confidence	Limits
therapy Drug vs Placebo			0.743	0.595	0.926
stratum 1 vs 4			0.203	0.145	0.282
stratum 2 vs 4			0.407	0.306	0.543
stratum 3 vs 4			0.534	0.398	0.716

Profile Likelihood Confidence Interval for Adjusted Odds Ratios					
95% Confidence Limits					
0.595			0.926		
0.144			0.281		
0.305			0.542		
0.397			0.715		

Output 1.17 lists the Wald chi-square statistics for the THERAPY and STRATUM variables, maximum likelihood estimates of the model parameters, associated odds ratios, and confidence intervals. The Wald statistic for the treatment effect equals 6.9635 and is significant at the 5% level ( $p = 0.0083$ ). This statistic is close in magnitude to the CMH statistic in Output 1.13. As shown by Day and Byar (1979), the CMH test is equivalent to the score test in a logistic regression model and is generally in a good agreement with the Wald statistic unless the data are sparse and most strata have only a few observations. Further, the maximum likelihood estimate of the overall ratio of the odds of survival is equal to 0.743, and the associated 95% Wald confidence interval is given by (0.595, 0.926). The estimate and confidence limits are consistent with the Mantel-Haenszel and logit-adjusted estimates and their confidence limits displayed in Output 1.13.

The bottom panel of Output 1.17 shows that the 95% profile likelihood confidence interval for the average odds ratio is equal to (0.595, 0.926). This confidence interval is requested by the CLODDS=PL option in the MODEL statement. Note that the profile likelihood confidence limits are identical to the Wald limits in this example. The advantage of using profile likelihood confidence intervals is that they are more stable than Wald confidence intervals in the analysis of very small or very large odds ratios. See Agresti (2002, Section 3.1).

Program 1.18 analyzes the same data set as above using PROC GENMOD. PROC GENMOD is a very flexible procedure for fitting various types of generalized linear models, including the logistic regression. The logistic regression model is requested by the DIST=BIN and LINK=LOGIT options. Alternatively, we can fit a model with a probit link function by setting DIST=BIN and LINK=PROBIT or any arbitrary link function (defined in the FWDLINK statement) that is consistent with the distribution of the response variable and desired interpretation of parameter estimates.

**PROGRAM 1.18 Maximum likelihood analysis of average association between treatment and survival in Case study 3 using PROC GENMOD**

```
proc genmod data=sepsis;
  class therapy stratum;
  model survival=therapy stratum/dist=bin link=logit type3;
  freq count;
  run;
```

**Output from Program 1.18**
**Analysis Of Parameter Estimates**

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits
Intercept		1	-0.1079	0.1082	-0.3199 0.1041
therapy	Drug	1	-0.2977	0.1128	-0.5189 -0.0766
therapy	Placebo	0	0.0000	0.0000	0.0000 0.0000
stratum	1	1	-1.5970	0.1695	-1.9292 -1.2648
stratum	2	1	-0.8981	0.1467	-1.1856 -0.6105
stratum	3	1	-0.6280	0.1499	-0.9217 -0.3343
stratum	4	0	0.0000	0.0000	0.0000 0.0000
Scale		0	1.0000	0.0000	1.0000 1.0000

**LR Statistics For Type 3 Analysis**

Source	DF	Chi-Square	Pr > ChiSq
therapy	1	6.99	0.0082
stratum	3	105.61	<.0001

Output 1.18 displays maximum likelihood estimates of the model parameters and the likelihood ratio test for the THERAPY variable. The computed estimates are different from those shown in Output 1.17 because PROC GENMOD relies on a different parameterization scheme for classification variables. The PROC GENMOD parameterization can be viewed as a more natural one since it is consistent with odds ratio estimates. For example, the maximum likelihood estimate of the overall ratio of the odds of survival (which is equal to 0.743) is easy to obtain by exponentiating the estimate of the treatment effect displayed in Output 1.18 (it equals  $-0.2977$ ). Further, if we are interested in computing the treatment effect estimate produced by PROC LOGISTIC ( $-0.1489$ ) and associated Wald statistic (6.9635), the PROC GENMOD code in Program 1.18 needs to be modified as follows:

```
proc genmod data=sepsis;
  class therapy stratum;
  model survival=therapy stratum/dist=bin link=logit type3;
  freq count;
  estimate "PROC LOGISTIC treatment effect" therapy 1 -1 /divisor=2;
  run;
```

Note that it is also possible to get PROC LOGISTIC to produce the PROC GENMOD estimates of model parameters displayed in Output 1.18. This can be done by adding the PARAM=GLM option to the CLASS statement as shown below

```

proc logistic data=sepsis;
  class therapy stratum/param=glm;
  model survival=therapy stratum/clodds=pl;
  freq count;
  run;

```

Again in Output 1.18, the likelihood ratio statistic for the null hypothesis of no treatment effect is requested by adding the TYPE3 option in the MODEL statement of PROC GENMOD. Output 1.18 shows that the statistic equals 6.99 and is thus close to the Wald statistic computed by PROC LOGISTIC. As pointed out by Agresti (2002, Section 5.2), the Wald and likelihood ratio tests produce similar results when the sample size is large. However, the likelihood ratio is generally preferred over the Wald test because of its power and stability.

Comparing the output generated by PROC GENMOD to that produced by PROC LOGISTIC, we can see that the latter procedure is more convenient to use because it computes odds ratios for each independent variable and associated confidence limits (both Wald and profile likelihood limits). Although the likelihood-ratio test for assessing the influence of each individual covariate on the outcome is not directly available in PROC LOGISTIC, this test can be carried out, if desired, by fitting two logistic models (with and without the covariate of interest) and then computing the difference in the model likelihood-ratio test statistics. Thus, PROC LOGISTIC matches all features supported by PROC GENMOD. One additional argument in favor of using PROC LOGISTIC in the analysis of stratified categorical data is that this procedure can perform exact inferences (in Version 8.1 and later versions of the SAS System) that are supported by neither PROC FREQ or PROC GENMOD. Exact tests available in PROC LOGISTIC are introduced in the next section.

### 1.3.5 Exact model-based tests

Exact inferences in PROC LOGISTIC are performed by conditioning on appropriate sufficient statistics. The resulting conditional maximum likelihood inference is generally similar to the regular (unconditional) maximum likelihood inference discussed above. The principal difference between the two likelihood frameworks is that the conditional approach enables us to evaluate the exact distribution of parameter estimates and test statistics and therefore construct exact confidence intervals and compute exact  $p$ -values. Mehta and Patel (1985); Stokes, Davis, and Koch (2000, Chapter 10); and Agresti (2002, Section 6.7) provided a detailed discussion of exact conditional inferences in logistic regression models.

Program 1.19 uses PROC LOGISTIC to conduct exact conditional analyses of the data set introduced in Section 1.3.2 (see Table 1.4). This data set contains mortality data collected at three centers in Case study 3. The EXACTONLY option in PROC LOGISTIC suppresses the regular output (asymptotic estimates and statistics), and the PARAM=REFERENCE option is added to the CLASS statement to allow the computation of exact confidence limits for the overall odds ratio. (The limits are computed only if the reference parameterization method is used to code classification variables.)

#### **PROGRAM 1.19    Exact conditional test of average association between treatment and survival at the three selected centers in Case study 3**

```

data sepsis1;
  input center therapy $ outcome $ count @@;
  if outcome="Dead" then survival=0; else survival=1;
  datalines;

```

```

1 Placebo Alive 2 1 Placebo Dead 2
1 Drug     Alive 4 1 Drug     Dead 0
2 Placebo Alive 1 2 Placebo Dead 2
2 Drug     Alive 3 2 Drug     Dead 1
3 Placebo Alive 3 3 Placebo Dead 2
3 Drug     Alive 3 3 Drug     Dead 0
;
proc logistic data=sepsis1 exactonly;
  class therapy center/param=reference;
  model survival(event="0")=therapy center;
  exact therapy/estimate=odds;
  freq count;
run;

```

---

**Output from  
Program 1.19**

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value ---	
			Exact	Mid
therapy	Score	4.6000	0.0721	0.0544
	Probability	0.0353	0.0721	0.0544

Exact Odds Ratios

Parameter	Estimate	95% Confidence		p-Value
		Limits		
therapy Drug	0.097	0.002	1.168	0.0751

---

Output 1.19 displays the exact  $p$ -values and confidence limits computed by PROC LOGISTIC. The exact  $p$ -values associated with the conditional score and probability methods are equal to the  $p$ -value produced by the Cochran-Armitage permutation test in PROC MULTTEST (see Output 1.14). The estimate of the average odds ratio of mortality in the experimental group compared to placebo equals 0.097 and lies in the middle between the Mantel-Haenszel and logit-adjusted estimates shown in Output 1.14 ( $\hat{\delta}_{MH} = 0.0548$  and  $\hat{\delta}_L = 0.1552$ ). The exact 95% confidence interval at the bottom of Output 1.19 is substantially wider than the asymptotic 95% confidence intervals associated with  $\hat{\delta}_{MH}$  or  $\hat{\delta}_L$ .

It is important to keep in mind that, up to Version 8.2 of the SAS System, the algorithm for exact calculations used in PROC LOGISTIC is rather slow. It may take several minutes to compute exact  $p$ -values in a data set with a thousand observations. In larger data sets, PROC LIFETEST often generates the following warning message:

“WARNING: Floating point overflow in the permutation distribution; exact statistics are not computed.”

Starting with Version 9.0, PROC LOGISTIC offers a more efficient algorithm for performing exact inferences and it becomes possible to run exact conditional analyses of binary outcomes even in large clinical trials. For example, it takes seconds to compute exact confidence limits and  $p$ -values in the trial data set in Case study 3 (see Program 1.20 below).

**PROGRAM 1.20    Exact conditional test of average association between treatment and survival in Case study 3**

```
proc logistic data=sepsis exactonly;
  class therapy stratum/param=reference;
  model survival(event="0")=therapy stratum;
  exact therapy/estimate=odds;
  freq count;
run;
```

**Output from  
Program 1.20**
**Exact Conditional Analysis**  
**Conditional Exact Tests**

Effect	Test	Statistic	--- p-Value ---	
			Exact	Mid
therapy	Score	6.9677	0.0097	0.0090
	Probability	0.00138	0.0097	0.0090

**Exact Odds Ratios**

Parameter	Estimate	95% Confidence		p-Value
		Limits		
therapy	Drug	0.743	0.592      0.932	0.0097

Output 1.20 lists the exact 95% confidence interval for the odds ratio of mortality (0.592, 0.932) and associated exact *p*-value (0.0097). The exact confidence limits are very close to the corresponding Mantel-Haenszel and logit-adjusted confidence limits shown in Output 1.13. Similarly, the exact *p*-value is in a good agreement with the CMH *p*-value for average association between treatment and survival (*p* = 0.0083).

Additionally, PROC LOGISTIC in Version 9.0 supports stratified conditional inferences proposed by Gail, Lubin, and Rubinstein (1981). The inferences are based on a logistic regression model with stratum-specific intercepts to better account for between-stratum variability. To request a stratified conditional analysis of the data in PROC LOGISTIC, we need to add the STRATA statement and specify the name of the stratification variable, e.g.,

```
proc logistic data=sepsis exactonly;
  class therapy/param=reference;
  model survival(event="0")=therapy;
  strata stratum;
  exact therapy/estimate=odds;
  freq count;
run;
```

An exact stratified conditional analysis of the SEPSIS data set generates confidence limits for the odds ratio of mortality and associated *p*-value that are identical to those displayed in Output 1.20.

### 1.3.6 Summary

This section reviewed statistical methods for the analysis of stratified categorical outcomes with emphasis on binary data. As was pointed out in the introduction, the analysis methods described in this section (e.g., Cochran-Mantel-Haenszel or model-based methods) are easily extended to a more general case of multinomial responses. PROC FREQ automatically switches to the appropriate extensions of binary estimation and testing procedures when it encounters outcome variables with three or more levels. Similarly, PROC LOGISTIC can be used to fit proportional-odds models for multinomial variables. See Chapter 3 of this book and Stokes, Davis and Koch (2000, Chapter 9) for details.

The first part of the section dealt with randomization-based estimates of the risk difference, relative risk and odds ratio and associated significance tests. The popular Cochran-Mantel-Haenszel test for overall association as well as the Mantel-Haenszel and logit-adjusted estimates of the average relative risk and odds ratio are easy to compute using PROC FREQ. The %MinRisk macro introduced in Section 1.3.3 implements minimum risk tests for association between treatment and a binary outcome in a stratified setting. The tests are attractive in clinical applications because they minimize the power loss under the worst possible configuration of stratum-specific parameters when patterns of treatment effects across strata are unknown.

Model-based tests for stratified binary data were discussed in the second part of the section. Model-based inferences can be implemented using PROC LOGISTIC and PROC GENMOD. PROC LOGISTIC appears to be more convenient to use in the analysis of stratified binary data because (unlike a more general PROC GENMOD) it generates a complete set of useful summary statistics for each independent variable in the model. For example, we can easily obtain odds ratios and associated Wald and profile likelihood confidence limits for each covariate. PROC LOGISTIC also supports exact inferences for stratified binary data.

When choosing an appropriate inferential method for stratified categorical data, it is important to remember that most of the popular procedures (both randomization- and model-based) need to be used with caution in sparse stratifications. The presence of a large number of under-represented strata either causes these procedures to break down or has a deleterious effect on their statistical power. A commonly used rule of thumb states that one generally needs at least five observations per treatment group per stratum to avoid spurious results. There is only a small number of exceptions to this rule. Both the Cochran-Mantel-Haenszel test and Mantel-Haenszel estimates of the average relative risk and odds ratio produced by PROC FREQ are known to be fairly robust with respect to stratum-specific sample sizes, and they perform well as long as the total sample size is large.

---

## 1.4 Analysis of time-to-event endpoints

This section reviews methods for randomization-based and model-based analysis in clinical trials with a time-to-event endpoint. Examples include survival endpoints, which are based on time to the onset of a therapeutic effect or time to worsening/relapse. First, we will discuss randomization-based tests for stratified time-to-event data and review stratified versions of the popular Wilcoxon and log-rank tests implemented in PROC LIFETEST as well as other testing procedures from the broad class of linear rank tests. The second part of this section covers model-based inferences for stratified time-to-event data that can be performed in the framework of the Cox proportional hazards regression. These inferences are

implemented using PROC PHREG. As in Section 1.3, this section deals only with fixed effects models for time-to-event outcomes. Random effects models will not be considered here. The reader interested in random effects models for time-to-event data used in clinical trials is referred to Andersen, Klein, and Zhang (1999) and Yamaguchi and Ohashi (1999).

**EXAMPLE: Case study 3 (Severe sepsis trial)**

In order to illustrate statistical methods for the analysis of time-to-event endpoints, we will consider an artificial data set containing 1600 survival times. The survival times are assumed to follow a Weibull distribution, i.e., the survival function in the  $i$ th treatment group and  $j$ th stratum is given by

$$S_{ij}(t) = \exp\{-(t/b_{ij})^a\},$$

where the shape parameter  $a$  equals 0.5, and the scale parameters  $b_{ij}$ 's shown in Table 1.5 are chosen in such a way that the generated survival times closely resemble the real survival times observed in the 1690-patient severe sepsis trial introduced in Section 1.3.

**TABLE 1.5 Scale parameters**

Stratum	Experimental group	Placebo group
1	$b_{11} = 13000$	$b_{21} = 25000$
2	$b_{12} = 13000$	$b_{22} = 10000$
3	$b_{13} = 5500$	$b_{23} = 3000$
4	$b_{14} = 2500$	$b_{24} = 1200$

The hazard function that specifies the instantaneous risk of death at time  $t$  conditional on survival to  $t$  is equal to  $h_{ij}(t) = at^{a-1}/b_{ij}^a$ . Since  $a$  equals 0.5, the hazard function is decreasing in a monotone fashion across all four strata.

Program 1.21 generates the SEPSURV data set with the SURVTIME variable capturing the time from the start of study drug administration to either patient's death or study completion measured in hours. The SURVTIME values are censored at 672 hours because mortality was monitored only during the first 28 days. The program uses PROC LIFETEST to produce the Kaplan-Meier estimates of survival functions across four strata. The strata were formed to account for the variability in the baseline risk of mortality. The TREAT variable in the SEPSURV data set identifies the treatment groups: TREAT=0 for placebo patients, and TREAT=1 for patients treated with the experimental drug.

**PROGRAM 1.21 Kaplan-Meier survival curves adjusted for the baseline risk of mortality**

```
data sepsurv;
  call streaminit(9544);
  do stratum=1 to 4;
    do patient=1 to 400;
      if patient<=200 then treat=0; else treat=1;
      if stratum=1 and treat=0 then b=25;
      if stratum=1 and treat=1 then b=13;
      if stratum=2 and treat=0 then b=10;
      if stratum=2 and treat=1 then b=13;
      if stratum=3 and treat=0 then b=3;
      if stratum=3 and treat=1 then b=5.5;
      if stratum=4 and treat=0 then b=1.2;
      if stratum=4 and treat=1 then b=2.5;
      survtime=rand("weibull",0.5,1000*b);
```

```

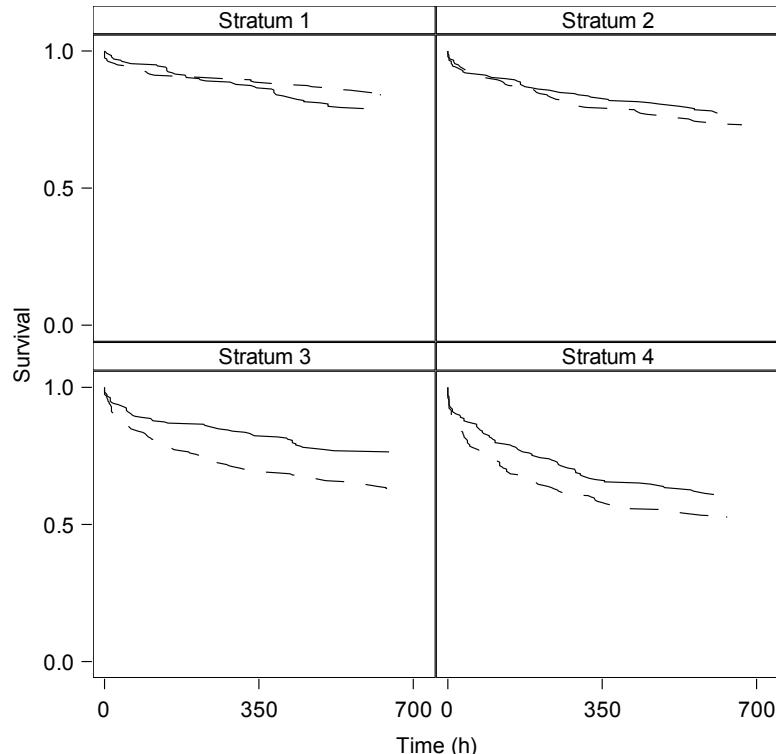
censor=(survtime<=672);
survtime=min(survtime,672);
output;
end;
end;

proc lifetest data=sepsurv notable outsurv=surv;
by stratum;
time survtime*censor(0);
strata treat;
run;

```

Figure 1.3 displays the Kaplan-Meier survival curves in the experimental and placebo arms across the strata representing increasing levels of mortality risk. It is clear that the placebo survival is significantly reduced in patients in Strata 3 and 4 compared to the experimental treatment. The beneficial effect of the experimental drug is most pronounced in patients at a high risk of death and the treatment effect is reversed in Stratum 1.

**Figure 1.3**  
Case study 3



*Kaplan-Meier survival curves adjusted for the baseline risk of mortality in the experimental arm (solid curve) and placebo arm (dashed curve).*

### 1.4.1 Randomization-based tests

In order to assess the significance of the treatment differences in Output 1.21 and test the null hypothesis of no treatment effect on survival in the 4 strata, we can make use of three randomization-based methods available in PROC LIFETEST: log-rank, Wilcoxon, and likelihood ratio tests.

The log-rank test was developed by Mantel (1966) who adapted the general Mantel-Haenszel approach (Mantel and Haenszel, 1959) to analyze right-censored, time-to-event data. Peto and Peto (1972) proposed a more general version of the original Mantel test for a comparison of multiple survival distributions. (Peto and Peto also coined the term “log-rank test.”) The Wilcoxon test in a two-sample scenario was developed by Gehan (1965). This test was later extended by Breslow (1970) to the case of multiple samples. Both the log-rank and Wilcoxon procedures are based on nonparametric ideas. The likelihood-ratio test is an example of a parametric method of comparing survival functions. The test is computed under the assumption that event times follow an exponential distribution. Alternatively, we can compare underlying survival distributions by carrying out generalized versions of *linear rank tests* described by Hájek and Šidák (1967). Examples include testing procedures proposed by Tarone and Ware (1977), Prentice (1978), and Harrington and Fleming (1982). The Tarone-Ware and Harrington-Fleming tests are closely related to the log-rank and Wilcoxon tests. In fact, the four testing procedures perform the same comparison of survival distributions but use four different weighting strategies. See Collett (1994, Section 2.5) and Lachin (2000, Section 9.3) for more details.

As an illustration, consider a clinical trial with a time-to-event endpoint comparing an experimental therapy to placebo. Let  $t_{(1)} < \dots < t_{(r)}$  denote  $r$  ordered event times in the pooled sample. The magnitude of the distance between two survival functions is measured in the log-rank test using the following statistic:

$$d_L = \sum_{k=1}^r (d_{1k} - e_{1k}),$$

where  $d_{1k}$  is the number of events observed in the experimental group at time  $t_{(k)}$  and  $e_{1k}$  is the expected number of events at time  $t_{(k)}$  under the null hypothesis of no treatment effect on survival. We can see from the definition of  $d_L$  that the deviations  $d_{1k} - e_{1k}$ ,  $k = 1, \dots, r$ , are equally weighted in the log-rank test.

The Wilcoxon test is based on the idea that early events are more informative than those that occurred later when few patients remain alive and survival curves are estimated with low precision. The Wilcoxon distance between two survival functions is given by

$$d_W = \sum_{k=1}^r n_k (d_{1k} - e_{1k}),$$

where  $n_k$  is the number of patients in the risk set before time  $t_{(k)}$ . (This includes patients who have not experienced the event of interest or have been censored before  $t_{(k)}$ .)

Similarly, the Tarone-Ware and Harrington-Fleming procedures are based on the following distance statistics:

$$d_{TW} = \sum_{k=1}^r n_k^{1/2} (d_{1k} - e_{1k}), \quad d_{HF} = \sum_{k=1}^r S_k^\rho (d_{1k} - e_{1k}),$$

respectively. Here  $S_k$  denotes the Kaplan-Meier estimate of the combined survival function of the two treatment groups at time  $t_{(k)}$ , and  $\rho$  is a parameter that determines how much weight is assigned to individual event times ( $0 \leq \rho \leq 1$ ). The weighting scheme used in the Tarone-Ware procedure represents the middle point between the case of equally-weighted event times and a Wilcoxon weighting scheme, which assigns greater weight to events that occurred early in the trial. To see this, note that the deviation  $d_{1k} - e_{1k}$  at time  $t_{(k)}$  receives the weight of  $n_k^0 = 1$  in the log-rank test and the weight of  $n_k^1 = n_k$  in the Wilcoxon test. The Harrington-Fleming procedure provides the statistician with a flexible balance between the log-rank-type

and Wilcoxon-type weights. Letting  $\rho = 0$  in the Harrington-Fleming procedure yields the log-rank test, whereas letting  $\rho = 1$  results in a test that assigns greater weights to early events.

An important consideration in selecting a statistical test is its efficiency against a particular alternative. The log-rank test is most powerful when the hazard functions in two treatment groups are proportional to each other but can be less efficient than the Wilcoxon test when the proportionality assumption is violated (Peto and Peto, 1972; Lee, Desu, and Gehan, 1975; and Prentice, 1978). It is generally difficult to characterize the alternative hypotheses that maximize the power of the Wilcoxon test because its efficiency depends on both survival and censoring distributions. It is known that the Wilcoxon test needs to be used with caution when early event times are heavily censored (Prentice and Marek, 1979). The Tarone-Ware procedure serves as a robust alternative to the log-rank and Wilcoxon procedures and maintains power better than these two procedures across a broad range of alternative hypotheses. Tarone and Ware (1977) demonstrated that their test is more powerful than the Wilcoxon test when the hazard functions are proportional, and it is more powerful than the log-rank test when the assumption of proportional hazards is not met. The same is true for the family of Harrington-Fleming tests.

## Comparison of survival distributions using the STRATA statement in PROC LIFETEST

Randomization-based tests for homogeneity of survival distributions across treatment groups can be carried out using PROC LIFETEST by including the treatment group variable in either the STRATA or TEST statement. For example, Program 1.22 examines stratum-specific survival functions in Case study 3. In order to request a comparison of the two treatment groups (experimental drug versus placebo) within each stratum, the TREAT variable is included in the STRATA statement.

### PROGRAM 1.22 Comparison of survival distributions in 4 strata using the STRATA statement

```
proc lifetest data=sepsurv notable;
  ods select HomTests;
  by stratum;
  time survtime*censor(0);
  strata treat;
  run;
```

---

#### Output from Program 1.22

The LIFETEST Procedure

##### Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	1.4797	1	0.2238
Wilcoxon	1.2748	1	0.2589
-2Log(LR)	1.6271	1	0.2021

stratum=2

## The LIFETEST Procedure

## Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.9934	1	0.3189
Wilcoxon	0.8690	1	0.3512
-2Log(LR)	1.1345	1	0.2868

stratum=3

## The LIFETEST Procedure

## Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	8.8176	1	0.0030
Wilcoxon	8.7611	1	0.0031
-2Log(LR)	10.3130	1	0.0013

stratum=4

## The LIFETEST Procedure

## Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	3.5259	1	0.0604
Wilcoxon	3.9377	1	0.0472
-2Log(LR)	4.4858	1	0.0342

Including the TREAT variable in the STRATA statement resulted in a comparison of stratum-specific survival distributions based on the log-rank, Wilcoxon, and likelihood ratio tests. Output 1.24 lists the log-rank, Wilcoxon, and likelihood ratio statistics accompanied by asymptotic *p*-values. Note that extraneous information has been deleted from the output using the ODS statement (`ods select HomTests`). We can see from Output 1.24 that the treatment difference is far from being significant in Stratum 1 and Stratum 2, highly significant in Stratum 3, and marginally significant in Stratum 4. The three tests yield similar results within each stratum with the likelihood ratio statistic consistently being larger than the other two. Since the likelihood-ratio test is a parametric procedure that relies heavily on the assumption of an underlying exponential distribution, it needs to be used with caution unless we are certain that the exponential assumption is met.

## Comparison of survival distributions using the TEST statement in PROC LIFETEST

An alternative approach to testing the significance of treatment effect on survival is based on the use of the TEST statement in PROC LIFETEST. If the treatment group variable is included in the TEST statement, PROC LIFETEST carries out only two tests (log-rank and Wilcoxon tests) to compare survival functions across

treatment groups. The generated log-rank and Wilcoxon statistics are somewhat different from those shown in Output 1.24.

As an illustration, Program 1.23 computes the stratum-specific log-rank and Wilcoxon statistics and associated *p*-values when the TREAT variable is included in the TEST statement.

### PROGRAM 1.23 Comparison of survival distributions in 4 strata using the TEST statement

```
proc lifetest data=sepsurv notable;
  ods select LogUniChisq WilUniChiSq;
  by stratum;
  time survtime*censor(0);
  test treat;
  run;
```

---

#### Output from Program 1.23

##### stratum=1

###### Univariate Chi-Squares for the Wilcoxon Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	-4.4090	3.9030	1.2761	0.2586

###### Univariate Chi-Squares for the Log-Rank Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	-5.2282	4.3003	1.4781	0.2241

##### stratum=2

###### Univariate Chi-Squares for the Wilcoxon Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	4.0723	4.3683	0.8691	0.3512

###### Univariate Chi-Squares for the Log-Rank Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	4.9539	4.9743	0.9918	0.3193

##### stratum=3

###### Univariate Chi-Squares for the Wilcoxon Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	13.8579	4.6821	8.7602	0.0031

###### Univariate Chi-Squares for the Log-Rank Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	16.2909	5.4914	8.8007	0.0030
<b>stratum=4</b>				
<b>Univariate Chi-Squares for the Wilcoxon Test</b>				
Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	10.3367	5.2107	3.9352	0.0473
<b>Univariate Chi-Squares for the Log-Rank Test</b>				
Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	12.3095	6.5680	3.5125	0.0609

Output 1.23 displays the Wilcoxon and log-rank statistics and *p*-values produced by Program 1.23. Note that the regular output also contains another set of Wilcoxon and log-rank statistics (under the headings “Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test” and “Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test”). These stepwise procedures are identical to the univariate procedures shown above because only one variable was included in the TEST statement in Program 1.23. The redundant test statistics and *p*-values were suppressed by using the ODS statement (`ods select LogUniChisq WilUniChiSq`). It is easy to check that the stratum-specific log-rank and Wilcoxon statistics in Output 1.23 are a bit different from those shown in Output 1.24. The difference is fairly small but tends to increase with the value of a test statistic. To understand the observed discrepancy between the two sets of test statistics, we need to remember that the TEST statement was added to PROC LIFETEST to enable the user to study effects of continuous covariates on survival distributions. The underlying testing procedures extend the log-rank and Wilcoxon tests for homogeneity of survival functions across treatment groups. See Kalbfleisch and Prentice (1980, Chapter 6). The more general versions of the log-rank and Wilcoxon tests do not always simplify to the ordinary tests when a categorical variable is included in the TEST statement. For example, if there were no identical survival times in the SEPSURV data set, the log-rank statistics in Output 1.23 would match those displayed in Output 1.24. Since this is not the case, the general version of the log-rank test implemented in PROC LIFETEST attempts to adjust for tied observations, which results in a slightly different value of the log-rank statistic. Analysis of time-to-event data with tied event times is discussed in more detail later in this section.

There is another way to look at the difference between inferences performed by PROC LIFETEST when the treatment group variable is included in the STRATA or TEST statement. Roughly speaking, the tests listed in Output 1.24 (using the STRATA statement) represent a randomization-based testing method, whereas the tests in Output 1.23 (using the TEST statement) are fairly closely related to a model-based approach. For example, it will be shown in Section 1.4.2 that the log-rank test in Output 1.23 is equivalent to a test based on the Cox proportional hazards model.

## Comparison of survival distributions using Tarone-Ware and Harrington-Fleming tests

Thus far, we have focused on the log-rank, Wilcoxon, and likelihood ratio testing procedures implemented in PROC LIFETEST. It is interesting to compare these procedures to the tests described by Tarone and Ware (1977) and Harrington and Fleming (1982). The latter tests can be carried out using PROC LIFETEST. For example, the following code can be used to compare survival distributions in the SEPSURV data set using the Tarone-Ware and Harrington-Fleming ( $\rho = 0.5$ ) tests.

The stratum-specific Tarone-Ware and Harrington-Fleming statistics and  $p$ -values produced by the %LinRank macro are summarized in Table 1.6. To facilitate the comparison with the log-rank and Wilcoxon tests computed in Program 1.24, the relevant test statistics and  $p$ -values from Output 1.24 are displayed at the bottom of the table.

**TABLE 1.6 Comparison of the Tarone-Ware, Harrington-Fleming, log-rank, and Wilcoxon tests**

Test	Stratum 1	Stratum 2	Stratum 3	Stratum 4
Tarone-Ware test				
Statistic	1.3790	0.9331	8.8171	3.7593
$P$ -value	0.2403	0.3341	0.0030	0.0525
Harrington-Fleming test ( $\rho = 0.1$ )				
Statistic	1.4599	0.9817	8.8220	3.5766
$P$ -value	0.2269	0.3218	0.0030	0.0586
Harrington-Fleming test ( $\rho = 1$ )				
Statistic	1.2748	0.8690	8.7611	3.9377
$P$ -value	0.2589	0.3512	0.0031	0.0472
Log-rank test				
Statistic	1.4797	0.9934	8.8176	3.5259
$P$ -value	0.2238	0.3189	0.0030	0.0604
Wilcoxon test				
Statistic	1.2748	0.8690	8.7611	3.9377
$P$ -value	0.2589	0.3512	0.0031	0.0472

**PROGRAM 1.24 Comparison of Survival distributions in Case study 3 using the Tarone-Ware and Harrington-Fleming tests**

```
proc lifetest data=sepsurv notable;
ods select HomTests;
by stratum;
time survtime*censor(0);
strata treat/test=(tarone fleming(0.5));
run;
```

Table 1.6 demonstrates that the  $p$ -values generated by the Tarone-Ware and Harrington-Fleming tests are comparable to the log-rank and Wilcoxon  $p$ -values computed in Program 1.24. By the definition of the Tarone-Ware procedure, the weights assigned to individual event times are greater than the log-rank weights and less than the Wilcoxon weights. As a consequence, the Tarone-Ware statistics lie between the corresponding log-rank and Wilcoxon statistics and the Tarone-Ware procedure is always superior to the least powerful of these two procedures. When  $\rho = 0.1$ , the Harrington-Fleming weights are virtually independent of event times. Thus, the stratum-specific Harrington-Fleming statistics are generally close in magnitude to the log-rank statistics. On the other hand, the Harrington-Fleming weights approximate the Wilcoxon weights when  $\rho = 1$ , which causes the two sets of test statistics to be very close to each other.

## Stratified analysis of time-to-event data

In the first part of this section, we have concentrated on stratum-specific inferences. However, we will typically be more interested in an overall analysis of the treatment effect that controls for important covariates. An adjusted effect of the experimental drug on a time-to-event outcome variable can be assessed by combining the information about the treatment differences across strata. PROC LIFETEST supports stratified versions of the Wilcoxon and log-rank tests. In order to carry out the stratified tests, we need to include the treatment group variable in the TEST statement and add the stratification variable (e.g., baseline risk of death) to the STRATA statement.

Mathematically, stratified inferences performed by PROC LIFETEST are equivalent to pooling the Wilcoxon and log-rank distance statistics with equal weights across  $m$  strata. Specifically, let  $d_{Lj}$  denote the value of the log-rank distance between survival functions in the  $j$ th stratum, and let  $s_{Lj}^2$  be the sample variance of  $d_{Lj}$ . The stratified log-rank statistic equals

$$u_L = \left( \sum_{j=1}^m d_{Lj} \right)^2 / \sum_{j=1}^m s_{Lj}^2.$$

Under the null hypothesis that there is no difference in underlying survival distributions, the stratified statistic asymptotically follows a chi-square distribution with 1 degree of freedom. The stratified Wilcoxon statistic is defined in a similar manner.

Program 1.25 conducts the two stratified tests to evaluate the significance of treatment effect on survival in Case study 3 adjusted for the baseline risk of mortality.

### **PROGRAM 1.25    Stratified comparison of survival distributions in 4 strata using the Wilcoxon and log-rank tests in Case study 3**

```
proc lifetest data=sepsurv notable;
  ods select LogUniChisq WilUniChiSq;
  time survtime*censor(0);
  strata stratum;
  test treat;
  run;
```

#### **Output from Program 1.25**

---

Univariate Chi-Squares for the Wilcoxon Test				
Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	23.8579	9.1317	6.8259	0.0090
Univariate Chi-Squares for the Log-Rank Test				
Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
treat	28.3261	10.7949	6.8855	0.0087

---

Output 1.25 shows the stratified Wilcoxon and log-rank statistics and  $p$ -values produced by Program 1.25. As in Program 1.23, the ODS statement is used (`ods select LogUniChisq WilUniChiSq`) to display the relevant sections of the PROC

LIFETEST output. The stratified Wilcoxon and log-rank  $p$ -values are 0.0090 and 0.0087, respectively. The two  $p$ -values are almost identical and indicate that the adjusted effect of the experimental drug on survival is very strong. It is worth noting that the stratified Wilcoxon and log-rank statistics are much larger than the corresponding statistics computed in Strata 1, 2 and 4 (see Output 1.23). As expected, stratification increases the power of the testing procedures.

We can also compare the computed stratified Wilcoxon and log-rank  $p$ -values to those produced by the stratified Tarone-Ware and Harrington-Fleming tests. From the %LinRank macro, the stratified Tarone-Ware statistic is 6.8975 ( $p = 0.0086$ ) and the stratified Harrington-Fleming statistic equals 6.9071 ( $p = 0.0086$ ) if  $\rho = 0.1$  and 6.8260 ( $p = 0.0090$ ) if  $\rho = 1$ . The obtained results are in a good agreement with the stratified analysis based on the Wilcoxon and log-rank tests.

### 1.4.2 Model-based tests

As in the case of categorical outcomes considered in Section 1.3, randomization-based tests for time-to-event data generally apply only to rather simple statistical problems, e.g., testing for equality of survival functions. Therefore, if we are interested in investigating effects of mixed (continuous and categorical) covariates on survival distributions or studying the prognostic ability of multiple factors, we need to rely on more flexible methods based on regression models for time-to-event data. Time-to-event outcome variables are analyzed using parametric models that explicitly specify the underlying survival distribution or semi-parametric models such as proportional hazards models. Parametric regression models are less commonly used in a clinical trial setting, and this section focuses on testing procedures related to proportional hazards regression. See Allison (1995, Chapter 4) and Cantor (1997, Chapter 5) for a review of parametric regression inferences based on PROC LIFEREG.

To define a proportional hazards model introduced in the celebrated paper by Cox (1972), consider a two-arm clinical trial with a time-to-event endpoint. The hazard function  $h_{ik}(t)$  for the  $k$ th patient in the  $i$ th treatment group is assumed to be

$$h_{ik}(t) = h_0(t) \exp\{X'_{ik}\beta_i\},$$

where  $h_0(t)$  denotes the unspecified baseline hazard function,  $X_{ik}$  is a covariate vector, and  $\beta_i$  is a vector of unknown regression parameters. The model is called a proportional hazards model because the ratio of the hazard functions in the two treatment groups is constant over time. Cox (1972) proposed to estimate the regression parameters by maximizing the *partial likelihood* that does not involve the unknown baseline hazard function  $h_0(t)$ . Although the partial likelihood is conceptually different from the ordinary or conditional likelihood (for instance, it lacks a probabilistic interpretation), the resulting estimates of the  $\beta_i$ 's possess the same asymptotic properties as the usual maximum likelihood estimates. As shown by Tsiatis (1981), maximum partial likelihood estimates are asymptotically consistent and normally distributed. For an extensive discussion of proportional hazards models, see Kalbfleisch and Prentice (1980, Chapters 4 and 5), Cox and Oakes (1984, Chapters 7 and 8), and Collett (1994, Chapter 3).

Proportional hazards models are implemented in PROC PHREG which supports a wide range of modeling features, (for example, stratified models, discrete-time models, and models with time-dependent covariates). Here, we will focus on proportional hazards models with time-independent covariates because clinical outcome variables are generally adjusted for baseline values of explanatory variables. See Allison (1995, Chapter 5) for examples of fitting proportional hazards models with time-dependent covariates in PROC PHREG.

In order to perform a stratified analysis of time-to-event data in PROC PHREG, the stratification variable needs to be specified in the STRATA statement. The STRATA statement enables us to fit a proportional hazards model when the hazard functions in the two treatment groups are parallel within each stratum but not across the entire sample. Program 1.26 fits a stratified proportional hazards model to the survival data collected in Case study 3 and plots the estimated survival functions. Note that PROC PHREG in Version 8 of the SAS System does not have a CLASS statement and thus one needs to create dummy indicator variables similar to the TREAT variable in the SEPSURV data set to include categorical covariates and interaction terms in the model. In order to eliminate this tedious pre-processing step, Version 9 has introduced an experimental version of PROC TPHREG with a CLASS statement.

**PROGRAM 1.26 Stratified comparison of survival distributions in 4 strata using the proportional hazards regression in Case study 3**

```

data cov;
    treat=0; output;
    treat=1; output;
proc phreg data=sepsurv;
    model survtime*censor(0)=treat/risklimits;
    strata stratum;
    baseline out=curve survival=est covariates=cov/nomean;

%macro PlotPH(stratum);
    axis1 minor=none label=(angle=90 "Survival") order=(0 to 1 by 0.5);
    axis2 minor=none label=("Time (h)") order=(0 to 700 by 350);
    symbol1 value=none color=black i=j line=1;
    symbol2 value=none color=black i=j line=20;
    data annotate;
        xsys="1"; ysys="1"; hsys="4"; x=50; y=20; position="5";
        size=1; text="Stratum &stratum"; function="label";
    proc gplot data=curve anno=annotate;
        where stratum=&stratum;
        plot est*survtime=treat/frame haxis=axis2 vaxis=axis1 nolegend;
        run;
        quit;
%mend PlotPH;

%PlotPH(1);
%PlotPH(2);
%PlotPH(3);
%PlotPH(4);

```

---

**Output from  
Program 1.26**

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.8861	1	0.0087
Score	6.8855	1	0.0087
Wald	6.8332	1	0.0089

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter	Standard	Chi-Square	Pr > ChiSq
		Estimate	Error		

treat	1	-0.24308	0.09299	6.8332	0.0089
<b>Analysis of Maximum Likelihood Estimates</b>					
Variable		Hazard Ratio	95% Hazard Ratio		
treat		0.784	0.654	0.941	

Output 1.26 displays the likelihood ratio, score and Wald statistics for testing the null hypothesis of no treatment effect. The three test statistics are very large in magnitude and indicate that the adjusted effect of the experimental drug on survival is highly significant. The score test is equivalent to the log-rank test for comparing survival distributions described earlier in Section 1.4.1. The two tests produce identical  $p$ -values when there are no tied event times (i.e., no two survival times are exactly the same). In the presence of tied observations, the log-rank is equivalent to the score test in PROC PHREG with a default adjustment for ties (Collett, 1994, Section 3.9). To see this, note that the adjusted log-rank statistic in Output 1.25 is equal to the score statistic in Output 1.26.

Output 1.26 presents qualitative information similar to that generated by PROC LIFETEST, but the lower panel describes the treatment difference in quantitative terms. Specifically, it shows the maximum partial likelihood estimate of the overall hazard ratio adjusted for the baseline risk of mortality and associated 95% confidence limits. The hazard ratio estimate equals 0.784, which means that the experimental drug reduces the hazard of death by 21.6%. The asymptotic 95% confidence interval for the true hazard ratio is equal to (0.654, 0.941). This confidence interval is requested by the RISKLIMITS option in the MODEL statement.

Figure 1.3 displays the estimated survival curves in the 4 strata from the CURVE data set generated by the BASELINE statement. In order to compute the survival curves in each of the two treatment groups, one needs to create a data set with two records TREAT=0 and TREAT=1 and pass the name of this data set to the COVARIATES option in the BASELINE statement. Since Program 1.26 performs a stratified analysis of the survival data, the shapes of the estimated survival curves vary from stratum to stratum but the treatment difference remains constant in the sense that the ratio of the hazard functions is the same in all 4 strata.

When fitting proportional hazards models in PROC PHREG, it is prudent to check whether the proportional hazards assumption is satisfied, i.e., the hazard ratio in any two treatment groups is approximately constant over time. Lagakos and Schoenfeld (1984) investigated the power of the score test when the assumption of proportional hazards is violated. They demonstrated that the score test is fairly robust with respect to modest deviations from the proportionality assumption, but its efficiency decreases dramatically if the hazard functions cross. Allison (1995, Chapter 5) discussed simple graphical checks of the proportional hazards assumption in PROC PHREG.

## Analysis of time-to-event data with ties

In general, analysis of discrete events of any kind becomes quite complicated in the presence of ties, i.e., events that occurred at exactly the same time. By running a frequency count of the SURVTIME variable, it is easy to check that there is a fair number of ties in the SEPSURV data set. For example, 4 patients died 10 hours after the start of study drug administration, and 5 patients died 15 hours after the start of study drug administration. Note that censored observations that occurred

at the same time are not counted as ties.

The presence of tied event times considerably complicates the derivation of the partial likelihood function and subsequent inferences. In order to streamline the inferences in data sets with ties, several authors have proposed to ignore the exact partial likelihood function and construct parameter estimates and significance tests based on an approximate partial likelihood function. Cox (1972) described an approximation to the partial likelihood function that assumes that the underlying time scale is discrete. Breslow (1974) proposed a simple approximation for the continuous-time case. The idea behind the Breslow method is that tied event times are treated as if there is a sequential ordering among them, and the likelihood function is averaged over all possible orderings. The Breslow method was improved by Efron (1977) who developed a more accurate approximation to the exact partial likelihood function. See Collett (1994, Section 3.3.2) for a good discussion of partial likelihood inferences in the presence of tied event times.

It is clear that different approaches to handling tied observations will yield different results and possibly lead to different conclusions. Since the method for handling ties was not explicitly specified in Program 1.26, PROC PHREG invoked the default method for handling ties developed by Breslow (1974). The Breslow method leads to biased parameter estimates in proportional hazards models when the number of ties is large. Since parameter estimates are biased toward 0, we are more likely to miss a significant effect when tied observations are accounted for inappropriately. For this reason, it is generally prudent to rely on more efficient adjustments for tied event times such as the Efron method or two exact methods available in PROC PHREG (EXACT and DISCRETE methods). The EXACT method uses the exact marginal likelihood function considered by Kalbfleisch and Prentice (1973). The DISCRETE method not only changes the method for handling tied observations but also changes the underlying model. Specifically, the TIES=DISCRETE option fits a discrete-time proportional odds model. Note that exact methods requested by the TIES=EXACT and TIES=DISCRETE options are more time-consuming than the Breslow or Efron methods. However, the difference is generally very small.

As an illustration, Program 1.27 fits a stratified proportional hazards model to the survival data collected in Case study 3 using the Efron and exact methods for handling ties.

**PROGRAM 1.27   Stratified comparison of survival distributions in 4 strata using the proportional hazards regression with the Efron and exact methods for handling ties in Case study 3**

```
title "Efron method for handling ties";
proc phreg data=sepsurv;
  model survtime*censor(0)=treat/ties=efron;
  strata stratum;
  run;
title "Exact method for handling ties";
proc phreg data=sepsurv;
  model survtime*censor(0)=treat/ties=exact;
  strata stratum;
  run;
```

---

**Output from  
Program 1.27**

Efron method for handling ties

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
------	------------	----	------------

Likelihood Ratio	6.9004	1	0.0086
Score	6.8998	1	0.0086
Wald	6.8475	1	0.0089

## Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter	Standard	Chi-Square	Pr > ChiSq
		Estimate	Error		
treat	1	-0.24333	0.09299	6.8475	0.0089

## Analysis of Maximum Likelihood Estimates

Variable	Hazard	95% Hazard Ratio	Confidence Limits
	Ratio		
treat	0.784	0.653	0.941

Exact method for handling ties

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.9004	1	0.0086
Score	6.8999	1	0.0086
Wald	6.8475	1	0.0089

## Analysis of Maximum Likelihood Estimates

Variable	Parameter	Standard	Chi-Square	Pr > ChiSq
	DF	Estimate		
treat	1	-0.24333	0.09299	6.8475

## Analysis of Maximum Likelihood Estimates

Variable	Hazard	95% Hazard Ratio	Confidence Limits
	Ratio		
treat	0.784	0.653	0.941

Output 1.27 lists the maximum partial likelihood estimates of the adjusted hazard ratio computed using the Efron and exact methods for handling tied survival times. The hazard ratio estimates as well as the 95% confidence limits are identical to those based on the default Breslow method in Output 1.26. Similarly, the likelihood ratio, score, and Wald statistics are only slightly different from the corresponding statistics in Output 1.26. Note that the score test in Output 1.27 is no longer equivalent to the stratified log-rank test because Program 1.27 makes use of the Efron and exact methods for handling ties.

### 1.4.3 Summary

This section reviewed randomization- and model-based testing procedures for stratified time-to-event data implemented in PROC LIFETEST and PROC PHREG. As always, randomization-based procedures can be used only for fairly simple inferences -

for example, for comparing survival distributions across treatment groups. Inferences based on regression models, such as the Cox proportional hazards regression, are more powerful. They enable the statistician to examine the effect of both continuous and categorical covariates on survival functions and to study the prognostic ability of multiple factors.

The following properties of randomization-based testing procedures need to be kept in mind when choosing a test for a particular time-to-event endpoint:

- The log-rank test is superior to other tests under the assumption of proportional hazards but is prone to losing power when this assumption is not satisfied.
- The Wilcoxon test is less sensitive to events occurring toward the end of a study when few patients remain alive.
- The Tarone-Ware and Harrington-Fleming procedures serve as a robust alternative to the log-rank and Wilcoxon procedures and maintain power better than these two procedures across a broad range of alternative hypotheses.

In general, given the complex nature of time-to-event data, it might be a good idea to estimate the power of candidate tests against the desirable alternatives via simulations.

Model-based tests for time-to-event data described in this section rely on the Cox proportional hazards regression with partial likelihood inferences. Proportional hazards models implemented in PROC PHREG exhibit several features that are very useful in applications:

- The baseline hazard function is completely removed from the inference. Therefore, the statistician does not need to specify the underlying survival distribution in order to estimate model parameters.
- Estimation in a proportional hazards model is performed nonparametrically (the inferences are based on ranked event times), which makes the method considerably more robust.
- PROC PHREG supports a wide range of useful features such as stratification, time-dependent covariates, and efficient adjustments for tied event times in proportional hazards models.

Inferences in proportional hazards models are generally fairly robust with respect to modest deviations from the proportionality assumption. However, a power loss is likely to occur when underlying hazards functions are not parallel to each other. When this happens, we can use the STRATA statement in PROC PHREG to fit separate proportional hazards models within each stratum.

## **1.5 Qualitative interaction tests**

---

The importance of an accurate assessment of the heterogeneity of the treatment effect among subgroups of interest has been emphasized in several publications and regulatory guidelines. For example, it is stated in Section 3.2 of the ICH guidance document “Statistical principles for clinical trials” (ICH E9) that

“If positive treatment effects are found in a trial with appreciable numbers of subjects per centre, there should generally be an exploration of the heterogeneity of treatment effects across centres, as this may affect the generalisability of the conclusions. Marked heterogeneity may be identified by graphical display of the

results of individual centres or by analytical methods, such as a significance test of the treatment-by-centre interaction.”

An appearance of reversed treatment differences in some strata does not always imply a true difference in the outcome variable across the chosen strata and might be due to random variation. In fact, we are virtually certain to encounter various degrees of inconsistency in the treatment effects across the strata. Senn (1997, Chapter 14) demonstrated that the probability of observing at least one effect reversal by chance increases rapidly with the number of strata. This probability exceeds 80% with 10 strata.

The following two definitions help differentiate between the cases of true and possibly random treatment-by-stratum interaction:

- A change in the magnitude of the treatment effect across the strata that does not affect the direction of the treatment difference is called a *quantitative* interaction. Quantitative interactions are very common in clinical trials. In fact, trials where the treatment difference is constant over all important patient subgroups are rarely seen. Quantitative interactions typically represent natural variability in the outcome variable and need to be contrasted with extreme cases involving qualitative interactions.
- A treatment-by-stratum interaction is termed *qualitative* if the direction of the true treatment difference varies across the strata. It is worth noting that qualitative interactions are sometimes referred to as *crossover* interactions.

Analysis of qualitative interactions have received a considerable amount of attention in the statistical literature (Peto, 1982; Azzalini and Cox, 1984; Gail and Simon, 1985; and Ciminera et al., 1993). The presence of qualitative interactions affects the interpretation of the overall trial outcome and influences the analysis of the treatment effect. Some authors, including Ciminera et al. (1993), advised against the use of quantitative interaction tests for judging the heterogeneity of the treatment effect and stressed that the treatment-by-stratum interaction needs to be included in the model only in the rare event of a pronounced qualitative interaction.

In this section, we will review the popular approach to testing for the presence of qualitative interactions proposed by Gail and Simon (1985) and briefly mention other qualitative interaction tests, including the test developed by Ciminera et al. (1993). Although the tests are often applied in the context of multicenter trials, it is clear that they can be used for the analysis of any stratification scheme.

### 1.5.1 Gail-Simon test

Gail and Simon (1985) proposed to formulate the problem of testing for qualitative interactions in terms of the following multivariate orthant hypotheses. Let  $\delta_i$  denote the true treatment difference, and let  $d_i$  and  $s_i$  denote the estimated treatment difference and associated standard error in the  $i$ th stratum. No qualitative interactions are present when the vector of the true treatment differences lies either in the positive orthant

$$O^+ = \{\delta_1 \geq 0, \dots, \delta_m \geq 0\}$$

or in the negative orthant

$$O^- = \{\delta_1 \leq 0, \dots, \delta_m \leq 0\}$$

of the  $m$ -dimensional parameter space. Gail and Simon described a likelihood ratio test for testing the null hypothesis of no qualitative interaction and demonstrated

that it can be expressed as

$$Q = \min(Q^+, Q^-) > c,$$

where  $Q^+$  and  $Q^-$  are given by

$$Q^+ = \sum_{i=1}^m \frac{d_i^2}{s_i^2} I(d_i > 0), \quad Q^- = \sum_{i=1}^m \frac{d_i^2}{s_i^2} I(d_i < 0)$$

and  $c$  is an appropriate critical value. The  $Q^+$  and  $Q^-$  statistics summarize the contribution of the positive and negative treatment differences, respectively. Gail and Simon also showed that the  $Q$  statistic follows a fairly complex distribution based on a weighted sum of chi-square distributions. This distribution can be used to derive the following formula for calculating the likelihood ratio  $p$ -value:

$$p = \sum_{i=1}^{m-1} (1 - F_i(Q)) \text{Bin}_{i,m-1}(0.5),$$

where  $F_i(x)$  is the cumulative distribution function of the chi-square distribution with  $i$  degrees of freedom and  $\text{Bin}_{i,m-1}(0.5)$  is the binomial probability mass function with success probability 0.5.

The described test is two-sided in the sense that the null hypothesis of no qualitative interaction is rejected when both  $Q^+$  and  $Q^-$  are large enough. The Gail-Simon method is easy to extend to one-sided settings to test (i) whether the true treatment differences are all positive or (ii) whether the true treatment differences are all negative. The  $p$ -value associated with the likelihood ratio test for (i) is equal to

$$p = \sum_{i=1}^m (1 - F_i(Q^-)) \text{Bin}_{i,m}(0.5)$$

and, similarly, the likelihood ratio  $p$ -value for testing (ii) is given by

$$p = \sum_{i=1}^m (1 - F_i(Q^+)) \text{Bin}_{i,m}(0.5).$$

## Gail-Simon test in Case study 1

We will first apply the introduced method to test for qualitative interaction in a clinical trial with a continuous endpoint. Program 1.28 carries out the introduced Gail-Simon test to examine the nature of the treatment-by-stratum interaction in Case study 1. The program calls the `%GailSimon` macro to compute the one- and two-sided likelihood ratio  $p$ -values. The macro has the following parameters:

- **Dataset** is the data set with test statistics and associated standard errors for each stratum.
- **Est** specifies the name of the variable containing the test statistics.
- **StdErr** specifies the name of the variable containing the standard errors.
- **TestType** is the type of the test to be carried out: one-sided Gail-Simon test for positive differences ("P"), one-sided Gail-Simon test for negative differences, or ("N") or two-sided Gail-Simon test ("T"), respectively.

**PROGRAM 1.28 Analysis of the qualitative interaction in Case study 1 using the Gail-Simon test**

```

proc sort data=hamd17;
  by drug center;
proc means data=hamd17 noprint;
  where drug="P";
  by center;
  var change;
  output out=plac mean=mp var=vp n=np;
proc means data=hamd17 noprint;
  where drug="D";
  by center;
  var change;
  output out=drug mean=md var=vd n=nd;
data comb;
  merge plac drug;
  by center;
  d=md-mp;
  n=np+nd;
  stderr=sqrt((1/np+1/nd)*((np-1)*vp+(nd-1)*vd)/(n-2));
%GailSimon(dataset=comb,est=d,stderr=stderr,testtype="P");
%GailSimon(dataset=comb,est=d,stderr=stderr,testtype="N");
%GailSimon(dataset=comb,est=d,stderr=stderr,testtype="T");

```

---

**Output from Program 1.28**

One-sided Gail-Simon test for positive differences

Test	
statistic	P-value
7.647	0.0465

One-sided Gail-Simon test for negative differences

Test	
statistic	P-value
57.750	0.0000

Two-sided Gail-Simon test

Test	
statistic	P-value
7.647	0.0297

---

We can see from Output 1.28 that all three likelihood ratio *p*-values are significant at the 5% level. The significance of the one-sided *p*-value for negative differences ( $Q^- = 57.750$ ,  $p < 0.0001$ ) is hardly surprising because it indicates that the true treatment differences are highly unlikely to be all negative. The other two tests are more informative and convey the same message —namely, the null hypothesis, that the positive treatment effect is consistent across the 5 selected centers, can be rejected. This means that the magnitude of the negative treatment difference at Center 101 is too large to be explained by chance and indicates the presence of qualitative interaction. Since Center 101 appears to be an outlier, it is prudent to carefully study the demographic and clinical characteristics of the patients enrolled at that site as well as consistency of the HAMD17 assessments to better understand the observed discrepancy.

### Gail-Simon test in Case study 3

Program 1.29 illustrates the use of the Gail-Simon test in the analysis of stratified categorical data. Consider the clinical trial in patients with severe sepsis introduced in Section 1.3. As shown in Figure 1.2, the treatment difference varies significantly across the 4 strata representing different levels of baseline risk of mortality. The survival rates are higher in patients treated with the experimental drug in Strata 2, 3, and 4. However, the treatment difference appears to be reversed in Stratum 1. Thus, it is reasonable to ask whether Stratum 1 is qualitatively different from the other strata. The program below uses the %GailSimon macro to test the consistency of the treatment differences in 28-day mortality across the 4 strata.

#### **PROGRAM 1.29 Analysis of the qualitative interaction in Case study 3 using the Gail-Simon test**

```

proc sort data=sepsis;
  by stratum;
data est;
  set sepsis;
  by stratum;
  retain ad dd ap dp;
  if therapy="Drug" and outcome="Alive" then ad=count;
  if therapy="Drug" and outcome="Dead" then dd=count;
  if therapy="Placebo" and outcome="Alive" then ap=count;
  if therapy="Placebo" and outcome="Dead" then dp=count;
  survd=ad/(ad+dd);
  survp=ap/(ap+dp);
  d=survrd-survp;
  stderr=sqrt(survrd*(1-survrd)/(ad+dd)+survp*(1-survp)/(ap+dp));
  if last.stratum=1;
%GailSimon(dataset=est,est=d,stderr=stderr,testtype="P");
%GailSimon(dataset=est,est=d,stderr=stderr,testtype="N");
%GailSimon(dataset=est,est=d,stderr=stderr,testtype="T");

```

---

#### **Output from Program 1.29**

One-sided Gail-Simon test for positive differences

Test	
statistic	P-value
0.855	0.6005

One-sided Gail-Simon test for negative differences

Test	
statistic	P-value
12.631	0.0030

Two-sided Gail-Simon test

Test	
statistic	P-value
0.855	0.4822

---

Both the one-sided statistic for positive differences ( $Q^+ = 0.855$ ) and the two-sided statistic ( $Q = 0.855$ ) in Output 1.29 indicate that there is little evidence in the data to reject the null hypothesis of no qualitative interaction. Despite the reversed treatment difference in Stratum 1, the Gail-Simon statistics are not large enough to conclude that this stratum is qualitatively different from the other strata at the 5% significance level. The highly non-significant  $p$ -values produced by the Gail-Simon test need to be contrasted with the Breslow-Day  $p$ -value displayed in Output 1.13. An application of the Breslow-Day test for quantitative homogeneity of odds ratios across the 4 strata in the same clinical trial yielded a  $p$ -value of 0.0899, which is clearly much closer to being significant than the conservative Gail-Simon  $p$ -values.

Finally, as in Output 1.28, the significant one-sided Gail-Simon statistic for negative differences ( $Q^- = 12.631$ ) suggests that the stratum-specific treatment differences in survival are very unlikely to be all negative.

### 1.5.2 Other qualitative interaction tests

Ciminera et al. (1993) described a test for qualitative interactions that serves as an alternative to the Gail-Simon test introduced earlier in this section. This test, known as the *pushback* test, is based on the following idea. To determine whether or not the observed treatment difference in a particular stratum is consistent with the rest of the data, we can order the standardized treatment differences and then compare with the expected values computed from the distribution of appropriately defined order statistics. This comparison will show which of the strata, if any, are more extreme than expected under the assumption of homogeneity. Ciminera et al. (1993) briefly compared the performance of the pushback and Gail-Simon tests and noted that the pushback procedure had fairly low power but was more sensitive than the Gail-Simon procedure in one example they studied.

### 1.5.3 Summary

This section discussed tests for qualitative interactions in clinical trials, including the popular Gail-Simon test. This test relies on stratum-specific summary statistics. Therefore, they can be carried out in clinical trials with continuous, categorical, or time-to-event endpoints.

As we pointed out in the introduction, tests for qualitative interaction have found many applications in clinical trials. These tests facilitate the analysis of treatment-by-stratum interactions, help identify subgroups of patients who have experienced the most pronounced beneficial effect, and play an important role in sensitivity analyses. For example, the trial's sponsor can carry out the Gail-Simon or other interaction tests to find one or several centers in a multicenter trial that are qualitatively different from the rest of the trial population. Although it may be difficult to justify an exclusion of these centers from the final analysis in the present regulatory environment, the obtained knowledge provides a basis for informative sensitivity analyses and ultimately leads to a better understanding of the trial results.

## 1.6 References

---

- Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine* 20, 2709-2722.
- Agresti, A. (2002). *Categorical Data Analysis* (Second Edition). New York: John Wiley and Sons.

- Agresti, A., Hartzel, J. (2000). Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine* 19, 1115-1139.
- Allison, P.D. (1995). *Survival Analysis Using the SAS System*. Cary, NC: SAS Institute, Inc.
- Andersen, P.K., Klein, J.P., Zhang, M. (1999). Testing for centre effects in multi-centre survival studies: A Monte Carlo comparison of fixed and random effects tests. *Statistics in Medicine* 18, 1489-1500.
- Azzalini, A., Cox, D.R. (1984). Two new tests associated with analysis of variance. *Journal of the Royal Statistical Society, Series B* 46, 335-343.
- Bancroft, T.A. (1968). *Topics in Intermediate Statistical Methods*. Ames, Iowa: Iowa University Press.
- Beach, M.L., Meier, P. (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials* 10, 161S-175S.
- Birch, M.W. (1964). The detection of partial association I. *Journal of the Royal Statistical Society, Series B* 26, 313-324.
- Breslow, N. (1970). A generalized Kruskal-Wallis test for comparing  $K$  samples subject to unequal patterns of censorship. *Biometrika* 57, 579-594.
- Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89-99.
- Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika* 68, 73-84.
- Breslow, N.E., Day, N.E. (1980). *Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies*. Lyon, Frances: International Agency for Research on Cancer: Lyon.
- Cantor, A. (1997). *Extending SAS Survival Analysis Techniques for Medical Research*. Cary, NC: SAS Institute, Inc.
- Chakravorti, S.R., Grizzle, J.E. (1975). Analysis of data from multiclinic experiments. *Biometrics* 31, 325-338.
- Chinchilli, V.M., Bortey, E.B. (1991). Testing for consistency in a single multi-center trial. *Journal of Biopharmaceutical Statistics* 1, 67-80.
- Ciminera, J.L., Heyse, J.F., Nguyen, H.H., Tukey, J.W. (1993). Tests for qualitative treatment-by-centre interaction using a “pushback” procedure. *Statistics in Medicine* 12, 1033-1045.
- Cochran, W.G. (1954a). The combination of estimates from different experiments. *Biometrics* 10, 101-129.
- Cochran, W.G. (1954b). Some methods for strengthening the common  $\chi^2$  tests. *Biometrics* 10, 417-451.
- Cochran, W.G. (1983). *Planning and Analysis of Observational Studies*. New York: Wiley.
- Coe, P.R., Tamhane, A.C. (1993). Small sample confidence intervals for the difference, ratio, and odds ratio of two success probabilities. *Communications in Statistics (Simulation and Computation)* 22, 925-938.
- Collett, D. (1994). *Modelling Survival Data in Medical Research*. London: Chapman and Hall.
- Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* 34, 187-220.
- Cox, D.R., Oakes, D.O. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

- Day, N.E., Byar, D.P. (1979). Testing hypotheses in case-control studies: Equivalence of Mantel-Haenszel statistics and logit score tests. *Biometrics* 35, 623-630.
- Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association* 72, 557-565.
- Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions* (Second edition). New York: Wiley.
- Fleiss, J.L. (1986). Analysis of data from multiclinic trials. *Controlled Clinical Trials* 7, 267-275.
- Ford, I., Norrie, J., Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine* 14, 735-746.
- Gail, M.H., Lubin, J.H., Rubinstein, L.V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* 68, 703-707.
- Gail, M., Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41, 361-372.
- Gail, M.H., Tan, W.Y., Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika* 75, 57-64.
- Gail, M.H., Wieand, S., Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71, 431-444.
- Gallo, P.P. (2000). Center-weighting issues in multicenter clinical trials. *Journal of Biopharmaceutical Statistics* 10, 145-163.
- Gastwirth, J.L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of the American Statistical Association* 80, 380-384.
- Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrary singly censored samples. *Biometrika* 52, 203-223.
- Goldberg, J.D., Koury, K.J. (1990). Design and analysis of multicenter trials. *Statistical Methodology in the Pharmaceutical Sciences*. Berry, D.A. (editor). New York: Marcel Dekker.
- Gould, A.L. (1998). Multi-center trial analysis revisited. *Statistics in Medicine* 17, 1779-1797.
- Greenland, S., Robins, J.M. (1985). Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 41, 55-68.
- Hájek, J., Šidák, Z. (1967). *Theory of Rank Tests*. New York: Academic Press.
- Hamilton M. (1967). Development of a rating scale for primary depressive illness. *British Journal of Clinical Psychology* 6, 278-296.
- Harrington, D.P., Fleming, T.R. (1982). A class of rank test procedures for censored survival data. *Biometrika* 69, 553-566.
- Hartley, H.O., Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* 54, 93-108.
- Hollander, M., Wolfe, D.A. (1999). *Nonparametric Statistical Method*. Second Edition. New York: Wiley.
- Hosmane, B., Shu, V., Morris, D. (1994). Finite sample properties of van Elteren test. *ASA Proceedings of the Biopharmaceutical Section* 430-434.
- Jones, B., Teather, D., Wang, J., Lewis, J.A. (1998). A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Statistics in Medicine* 17, 1767-1777.

- Kalbfleisch, J.D., Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* 60, 267-278.
- Kalbfleisch, J.D., Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Kallen, A. (1997). Treatment-by-center interaction: What is the issue? *Drug Information Journal* 31, 927-936.
- Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E. (1985). APACHE II: A severity of disease classification system. *Critical Care Medicine* 13, 818-829.
- Koch, G.G., Amara, I.A., Davis, G.W., Gillings, D.B. (1982). A review of some statistical methods for covariance analysis of categorical data. *Biometrics* 38, 563-595.
- Koch, G.G., Gillings, D.B. (1983). Inference, design-based vs. model-based. *Encyclopedia of Statistical Sciences*. Kotz, S., Johnson, N.L., Read, C.B. (editors). New York: Wiley.
- Koch, G.G., Carr, G.J., Amara, I.A., Stokes, M.E., Uryniak, T.J. (1990). Categorical data analysis. *Statistical Methodology in the Pharmaceutical Sciences*. Berry, D.A. (editor). New York: Marcel Dekker.
- Koch, G.G., Edwards, S. (1988). Clinical efficacy with categorical data. *Biopharmaceutical Statistics for Drug Development*. Peace, K.K. (editor). New York: Marcel Dekker.
- Lachin, J.M. (2000). *Biostatistical Methods. The Assessment of Relative Risks*. New York: Wiley.
- Lagakos, S.W., Schoenfeld, D.A. (1984). Properties of proportional hazards score tests under misspecified regression models. *Biometrics* 40, 1037-1048.
- Lee, E.T., Desu, M.M., Gehan, E.A. (1975). A Monte Carlo study of the power of some two-sample tests. *Biometrika* 62, 425-432.
- Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco, CA: Holden-Day.
- Lin, Z. (1999). An issue of statistical analysis in controlled multicentre studies: How shall we weight the centers? *Statistics in Medicine* 18, 365-373.
- Littell, R.C., Freund, R.J., Spector, P.C. (1991). *SAS System for Linear Models* (Third Edition). Cary, NC: SAS Institute, Inc.
- Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institute, Inc.
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo Reports* 50, 163-170.
- Mantel, N., Fleiss, J.L. (1980). Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure. *American Journal of Epidemiology* 112, 129-134.
- Mantel, N., Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22, 719-748.
- Mehrotra, D.V. (2001). Stratification issues with binary endpoints. *Drug Information Journal* 35, 1343-1350.
- Mehrotra, D.V., Railkar, R. (2000). Minimum risk weights for comparing treatments in stratified binomial trials. *Statistics in Medicine* 19, 811-825.
- Mehta, C.R., Patel, N.R. (1985). Exact logistic regression: theory and examples. *Statistics in Medicine* 14, 2143-2160.
- Milliken, G.A., Johnson, D.E. (1984). *Analysis of Messy Data: Designed Experiments*. London: Chapman and Hall.

- Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society, Series A* 140, 48-76.
- Nelder, J.A. (1994). The statistics of linear models: back to basics. *Statistics and Computing* 4, 221-234.
- Peto, R. (1982). Statistical aspects of cancer trials. *Treatment of Cancer*. Halnan, K.E. (editor). London: Chapman and Hall.
- Peto, R., Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A* 135, 185-206.
- Phillips, A., Ebbutt, A., France, L., Morgan, D. (2000). The International Conference on Harmonization guideline “Statistical principles for clinical trials”: issues in applying the guideline in practice. *Drug Information Journal* 34, 337-348.
- Prentice, R.L. (1978). Linear rank tests with right censored data. *Biometrika* 65, 167-180.
- Prentice, R.L., Marek, P. (1979). A qualitative discrepancy between censored data rank. *Biometrics* 35, 861-867.
- Radhakrishna, S. (1965). Combination of results from several  $2 \times 2$  contingency tables. *Biometrics* 21, 86-98.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Application* (Second Edition). New York: Wiley.
- Robinson, L.D., Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 58, 227-240.
- Rodriguez, R., Tobias, R., Wolfinger, R. (1995). Comments on J.A. Nelder ‘The Statistics of Linear Models: Back to Basics’. *Statistics and Computing* 5, 97-101.
- Scheffe, H. (1959). *The Analysis of Variance*. New York: Wiley.
- Searle, S.R. (1971). *Linear Models*. New York: Wiley.
- Searle, S.R. (1976). Comments on ANOVA calculations for messy data. *SAS Users Group International Conference I* 298-308.
- Searle, S.R. (1992). *Variance Components*. New York: Wiley.
- Senn, S. (1997). *Statistical Issues in Drug Development*. New York: Wiley.
- Senn, S. (1998). Some controversies in planning and analyzing multicenter trials. *Statistics in Medicine* 17, 1753-1765.
- Senn, S. (2000). The many modes of meta. *Drug Information Journal* 34, 535-549.
- Speed, F.M., Hocking, R.R. (1976). The use of  $R()$  notation with unbalanced data. *American Statistician* 30, 30-33.
- Speed, F.M., Hocking, R.R. (1980). A characterization of the GLM sums of squares. *SAS Users Group International Conference V* 215-223.
- Speed, F.M., Hocking, R.R., Hackney, O.P. (1978). Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association* 73, 105-112.
- Stokes, M.E., Davis, C.S., Koch, G.G. (2000). *Categorical Data Analysis Using the SAS System* (Second edition). Cary, NC: SAS Institute, Inc.
- Tarone, R.E., Ware, J. (1977). On distribution-free tests for equality of survival distribution. *Biometrika* 64, 156-160.
- Tsiatis, A.A. (1981). A large sample study of Cox’s regression model. *Annals of Statistics* 9, 93-108.
- van Elteren, P.H. (1960). On the combination of independent two-sample tests of Wilcoxon. *Bulletin of the International Statistical Institute* 37, 351-361.

Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., Hochberg, Y. (1999).  
*Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Institute, Inc.

Yamaguchi, T., Ohashi, Y. (1999). Investigating centre effects in a multi-centre clinical trial of superficial bladder cancer. *Statistics in Medicine* 18:1961-1971.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association* 29, 51-66.

# Chapter 2

## Advanced Randomization-based Methods

**Richard C. Zink (JMP Life Sciences, SAS Institute)**

**Gary G. Koch (University of North Carolina at Chapel Hill)**

**Yunro Chung (Fred Hutchinson Cancer Research Center)**

**Laura Elizabeth Wiener (University of North Carolina at Chapel Hill)**

2.1	Introduction	67
2.2	Case studies	70
2.3	%NParCov4 macro	73
2.4	Analysis of ordinal endpoints using a linear model	74
2.5	Analysis of binary endpoints	78
2.6	Analysis of ordinal endpoints using a proportional odds model	79
2.7	Analysis of continuous endpoints using the log-ratio of two means	80
2.8	Analysis of count endpoints using log-incidence density ratios	81
2.9	Analysis of time-to-event endpoints	82
2.10	Summary	86

This chapter presents advanced randomization-based methods used in the analysis of clinical endpoints. The nonparametric, randomization-based analysis of covariance methodology, a versatile methodology that uses weighted least squares to generate covariate-adjusted treatment effects under minimal assumptions, is presented. These methods are implemented using a powerful SAS macro (%NParCov4) that is applicable to a variety of clinical outcomes with the ability to compute essentially-exact  $p$ -values and confidence intervals. Numerous clinical trial examples are presented to illustrate the methodology and software.

### 2.1 Introduction

---

As explained in Chapter 1, adjustments for important covariates in the analysis of clinical endpoints can be carried out using randomization-based and model-based methods. The two classes of methods play complementary roles in the sense that each method is useful within a particular setting, and the advantages of each method offset the limitations of the other method. Chapter 1 deals with both the model-based and basic randomization-based methods applied in clinical trials. This chapter focuses on a class of advanced randomization-based methods with broad applications to trials with different types of endpoints.

A detailed description of model-based approaches can be found in the beginning of Chapter 1. This includes, for example, logistic regression models used in the analysis of binary endpoints and the Cox proportional hazards model in settings with time-to-event endpoints. An important feature of model-based methods is that they rely on specific assumptions and thus can be quite sensitive to violations from these assumptions. For example, considering proportional hazards models, departures from the proportional hazards assumption are typically more evident for baseline characteristics and measurements than for treatments. Also, proportional hazards is applicable to treatments under the null hypothesis of equal treatment effects. The issue of bias in a treatment comparison through departures from the proportional hazards assumption for baseline characteristics and/or measurements can be addressed by a randomization-based method that invokes randomization-based covariance adjustment for the baseline characteristics and/or measurements for an estimated hazard ratio from a proportional hazards model that only includes treatments; see Saville and Koch (2013). Similar considerations apply to the proportional odds model with respect to how to address departures from the proportional odds assumption for baseline characteristics and/or measurements. See Tangen and Koch (1999a). Additionally, the scope of randomization-based methods for ordinal response includes Mann-Whitney rank measures of association. See Kawaguchi and Koch (2011, 2015).

---

## **Nonparametric randomization-based analysis of covariance**

---

It was pointed out in Chapter 1 that analysis of covariance serves two important purposes in a randomized clinical trial. First, there is a reduction of variance for the treatment estimate, which provides a more powerful statistical test and a more precise confidence interval. Second, analysis of covariance provides an estimate of the treatment effect, which is adjusted for random imbalances of covariates between the treatment groups. Additionally, analysis of covariance models provide estimates of covariate effects on the outcome (adjusted for treatment) and can test for the homogeneity of the treatment effect within covariate-defined subgroups. The nonparametric, randomization-based analysis of covariance method of Koch et al. (1998) defines a versatile methodology that uses weighted least squares to generate covariate-adjusted treatment effects with minimal assumptions. This methodology is general in its applicability to a variety of outcomes, whether continuous, binary, ordinal, incidence density, or time-to-event. (See Koch et al., 1998; Tangen and Koch, 1999a, 1999b, 2000; LaVange and Koch, 2008; Moodie et al., 2010; Saville, Herring, and Koch, 2010; Saville, LaVange, and Koch, 2010; Saville and Koch 2013; and Zhao et al. 2016.) Further, its use has been illustrated in many clinical trial settings, such as multi-center, dose-response, and non-inferiority trials (Koch and Tangen, 1999; LaVange, Durham and Koch, 2005; and Tangen and Koch, 2001). Similar covariate adjustment for randomized studies has been applied to multivariate Mann-Whitney estimators of ordinal outcomes (Kawaguchi, Koch, and Wang, 2011; and Kawaguchi and Koch, 2015).

The nonparametric, randomization-based analysis of covariance methodology uses weighted least squares on the treatment differences of outcome and covariate means. The resulting model estimates represent treatment effects for the outcomes that have been adjusted for the accompanying covariates. Covariance-adjustment in the treatment-effect estimates is a result of the assumed null difference in covariate means, which, in turn, is a consequence of the underlying invocation of randomization to treatment. Full details of this methodology are left to Appendix A.

This methodology has several advantages:

1. Applicability to a variety of outcomes, whether continuous, binary, ordinal, incidence density, or time-to-event (See Appendices A.1 through A.5).
2. Minimal assumptions:
  - (a) Randomization to treatment is the principal structure necessary for hypothesis testing.
  - (b) Confidence intervals of treatment estimates require the additional assumption that subjects in the trial are a simple random sample of a larger population.
  - (c) Survival analyses have the assumption of non-informative censoring.
3. Straightforward to accommodate stratification and to generate essentially-exact  $p$ -values and confidence intervals (described in Appendices A.6 and A.7, respectively).
4. Greater power of the adjusted treatment effect relative to the unadjusted.

However, there are some disadvantages:

1. No estimates for the effects of covariates or strata are available.
2. No direct way to test homogeneity of treatment effects within subgroups that are defined by the covariates.
3. Results are not directly generalizable to subjects with covariate distributions that differ from those in the trial.
4. Is not appropriate when the same multivariate distribution of the covariates cannot be assumed to apply to both treatment groups.

It is important to note that nonparametric, randomization-based analysis of covariance is not intended to replace likelihood-based methodologies such as generalized linear models (see Chapter 1). On the contrary, this methodology is expected to be used in a complementary fashion with model-based methods, as long as patients are randomized to treatment (see Tangen and Koch, 1999a, 2000; and Saville and Koch, 2013, as examples). In a regulatory setting, randomization-based, nonparametric analysis of covariance can establish whether a treatment effect exists under minimal statistical assumptions. Once a treatment effect has been identified, other scientific questions can be explored under some additional plausible assumptions using more traditional models. The examples in this chapter focus on the analysis and interpretation of data using nonparametric, randomization-based analysis of covariance. Additional analyses to address other scientific questions are not shown. (Interested readers might choose to use the model-based methods introduced in Chapter 1 to compare with the findings presented here.)

For example, as outlined in Tangen and Koch (1999a), the specification of statistical methods in a protocol or analysis plan is typically required prior to the unblinding of study data. It is recommended that randomization-based analysis of covariance be specified as the primary analysis because of minimal assumptions and the availability of essentially-exact  $p$ -values. Model-based methods, such as a logistic regression model in the case of binary outcomes, can be used in a supportive fashion to assess the effects of covariates, potential interactions of treatment with covariates, and the generalizability of findings to populations not represented by the subjects in the current study. This last point is particularly important. For nonlinear models, such as logistic regression, the estimated treatment effect is specific to a population of patients with a specific set of values for the covariates. Randomization-based, nonparametric analysis of covariance estimates and tests the treatment effect with adjustment for covariates for the randomized population of patients.

This chapter provides a detailed description of the nonparametric, randomization-based analysis of covariance methodology. A description of key features of the advanced randomization-based framework is provided in Section 2.1. Case studies used throughout this chapter are presented in Section 2.2. Section 2.3 introduces the `%NParCov4` macro that performs these analyses with the ability to compute essentially exact  $p$ -values and confidence intervals. We discuss the strengths and limitations of the randomization-based methodology and provide complementary analysis strategies using the case studies in Sections 2.4 through 2.9. As stated above, technical details are left to the Appendix. Appendix A summarizes the theoretical details of these methods while Appendix B documents the `%NParCov4` macro.

The SAS code and data sets included in this chapter are available on the book's website at <http://support.sas.com/publishing/authors/dmitrienko.html>.

## 2.2 Case studies

---

The following four case studies, based on real clinical trials, will be used in this chapter to illustrate advanced randomization-based methods for the analysis of clinical endpoints.

### 2.2.1 Case study 1 (Respiratory disorder trial)

The goal of this randomized clinical trial was to compare an active treatment versus placebo for 111 subjects with a respiratory disorder at two centers (Koch et al., 1990; and Stokes, Davis, and Koch, 2012). The primary endpoint was ordinal, and subjects were classified with an overall score of 0, 1, 2, 3, or 4 (terrible, poor, fair, good, or excellent) at each of four visits (V1 through V4). Covariates included a baseline score, gender, and age (in years) at study entry. Treatment was coded 1 for active treatment, 0 for placebo, while gender was coded 1 for male, and 0 for female. In addition to the ordinal scale, dichotomous outcomes, such as (good, excellent) versus (terrible, poor, fair) with values of 1 or 0, respectively, were of interest. The first few rows of the respiratory data are presented in Table 2.1.

**TABLE 2.1** Respiratory disorder data

Center	Treatment	Gender	Age	Baseline	V1	V2	V3	V4
0	0	1	46	2	2	2	2	2
0	0	1	28	2	0	0	0	0
0	0	1	44	3	4	3	4	2
0	0	0	13	4	4	4	4	4
0	0	1	43	2	3	2	4	4
0	0	1	37	3	2	3	3	2
0	0	1	23	3	3	1	1	1
:	:	:	:	:	:	:	:	:

There are several ways to analyze the overall score from this clinical trial. For example, if the distribution of overall scores at a visit appears to have an approximately normal shape, a linear model might be appropriate to assess the effect of treatment with adjustment for covariates (Section 2.4). Alternatively, the clinical team might prefer to compare the proportion of patients with good or excellent response between the two treatments. Here, a logistic regression model can be used to model the effect of covariates on treatment response (Section 2.5). Finally, if the proportional odds assumption is appropriate, a proportional odds

model can be fit to the overall score with adjustment for covariates (Section 2.6). The proportional odds assumption implies that the model coefficients for all binary splits of the outcome (0 vs >0,  $\leq 1$  vs >1,  $\leq 2$  vs >2, and  $\leq 3$  vs 4) are identical.

### 2.2.2 Case study 2 (Osteoporosis trial)

Data from a randomized clinical trial in 214 postmenopausal women with osteoporosis at two study centers were used to evaluate the effectiveness of a new drug (t) versus placebo (p) (Stokes et al., 2012). Both treatment groups had access to calcium supplements, received nutritional counseling, and were encouraged to exercise regularly. The number of fractures experienced each year were recorded through the trial's duration (three years). Seven women discontinued during the third year and are considered to have a 6-month exposure for this year. Otherwise, yearly exposure is considered to be 12 months. The outcome of interest was the rate of fracture per year. Adjusting for age as a covariate and for centers as strata would be appropriate.

Table 2.2 contains the first few rows of the osteoporosis data set. Year 1, Year 2, and Year 3 contain the number of fractures experienced in Years 1, 2, and 3, respectively. Total is the total number of fractures experienced over the duration of the trial. Exposure equals 3 years save for the 7 patients who discontinued early with 2.5 years of exposure. Yearly risk is defined as the total number of fractures (Total) divided by Exposure. Center A and B were recoded as 1 and 2, while Treatment p and t were recoded as 1 and 2 for the analysis. Yearly risk can be analyzed using a Poisson regression model to assess the effect of treatment with adjustment for covariates (Section 2.7).

**TABLE 2.2** Osteoporosis data

Age	Center	Treatment	Year 1	Year 2	Year 3	Total	Exposure
56	A	p	0	0	0	0	3
71	A	p	1	0	0	1	3
60	A	p	0	0	1	1	3
71	A	p	0	1	0	1	3
78	A	p	0	0	0	0	3
67	A	p	0	0	0	0	3
49	A	p	0	0	0	0	3
:	:	:	:	:	:	:	:

### 2.2.3 Case study 3 (Aneurysmal subarachnoid hemorrhage trial)

Nicardipine is indicated for the treatment of high blood pressure and angina in both oral and intravenous formulations. Data from 902 subjects from a randomized, two-arm intravenous study of high dose nicardipine was presented in Haley, Kassell, and Torner (1993). The primary endpoint was improvement in patient recovery according to the Glasgow Outcome Scale, with the incidence of cerebral vasospasm and the incidence of death or disability due to vasospasm serving as important secondary endpoints (Jennett and Bond, 1975). Subjects were treated with drug or placebo up to 14 days after the hemorrhage occurred. Here, we compare the number of treatment-emergent, serious adverse events between nicardipine (Treatment = 1) and placebo (Treatment = 0), accounting for the varying durations of treatment exposure. Age in years and gender are available as covariates.

Table 2.3 contains the first few rows of the aneurysmal, subarachnoid hemorrhage data set. Count is the total number of treatment-emergent, serious adverse events

that occur over the duration of trial exposure (Days). Male is coded 1 if the patient is a male, and 0 if the patient is a female. The number of adverse events can be analyzed using a Poisson regression model using exposure as an offset to assess the effect of treatment with adjustment for covariates (Section 2.8).

**TABLE 2.3 Aneurysmal, subarachnoid hemorrhage Data**

Treatment	Count	Days	Age	Male
0	0	13	63	0
1	0	11	66	1
0	0	10	31	0
1	6	1	48	0
0	1	13	67	0
0	0	11	32	1
1	5	3	63	1
:	:	:	:	:

## 2.2.4 Case study 4 (Colorectal cancer trial)

Data from a randomized study of 110 subjects with colorectal cancer are presented to compare an experimental treatment to a control (Table 2.4). Treatment (Trt) is coded 1 or 0 for active treatment or control, respectively. Data for two endpoints are available: the time until death and the time until disease progression. Disease progression was defined using standardized RECIST criteria (Eisenhauer et al., 2009). Subjects who did not experience death or disease progression were censored on their last day of study participation. Overall survival (OS) and progression-free survival (PFS) are each represented by two variables. The variable OS\_Time is the number of days from randomization until death. OS\_Event equals 1 if the patient died. Otherwise, a value of 0 indicates that the patient was censored on their last day of study participation. For PFS, the variable PFS\_Time is the number of days from randomization until disease progression. PFS\_Event equals 1 if the patient experienced disease progression. Otherwise, a value of 0 indicates that the patient was censored on their last day of study participation.

**TABLE 2.4 Colorectal cancer data**

Trt	OS_Time	OS_Event	PFS_Time	PFS_Event	KRAS	Days	Length
1	75	1	40	1	WT	805	121
1	406	0	68	1	WT	3186	160
1	225	1	118	1	M	2449	101
1	89	1	36	1	M	457	97
1	309	1	72	1	WT	1396	158
0	123	1	50	1	WT	2341	120
1	465	0	48	1	WT	1513	64
:	:	:	:	:	:	:	:

Covariates include the number of days from initial cancer diagnosis to start of study drug administration (Days), and the sum of the longest lesion diameters in millimeters at baseline (Length). A flag as to whether the KRAS gene was wild-type (WT) or mutated (M) can be used as a stratification variable. For the analysis, a numeric variable mutant will be defined with values of 0 or 1 to indicate wild-type or mutated, respectively. Time-to-event endpoints are often analyzed using non-parametric Kaplan-Meier analysis (Section 2.9). Alternatively, Cox proportional-hazards regression is used when it is important to adjust for one or more covariates.

## 2.3 %NParCov4 macro

---

Analysis of clinical endpoints based on the nonparametric, randomization-based methodology is easily performed using the %NParCov4 macro, a SAS/IML macro that supports functionality for many of the analysis methods described to date, with the ability to compute essentially-exact *p*-values and confidence intervals. This macro will be used in Sections 2.4 through 2.9 to perform adjusted analyses of continuous, categorical, count-type, and time-to-event endpoints.

A macro call including all options is presented in Program 2.1.

### PROGRAM 2.1 General %NParCov4 Macro Call

```
%NPARCOV4(OUTCOMES = , COVARS = , EXPOSURES = , TRTGRPS = , HYPOTH = ,
TRANSFORM = , STRATA = , COMBINE = , C = , ALPHA = , EXACT = , SEED = ,
NREPS = , SYMSIZE = , DETAILS = , DSNIN = , DSNOUT = );
```

The user can specify one or more numeric outcomes (**OUTCOMES**), one or more numeric covariates, or none if no covariate adjustment is required (**COVARS**). For time-to-event outcomes or for analyses of incidence densities, the user specifies one or more numeric exposure variables (**EXPOSURES**). A two-level numeric variable is required to define treatment (**TRTGRPS**).

Hypothesis (**HYPOTH**) is specified as either **NULL** or **ALT** and impacts whether the covariance matrices of responses and covariates in each strata are computed across treatments (under the null) or within treatments (under the alternative). **HYPOTH = NULL** by default. In general, the variance under the null would be the principal structure for producing *p*-values for hypothesis testing under the null when the null corresponded to no difference between treatments (or more formally, the strong null that each patient would have the same outcomes regardless of the assigned treatment). The variance under the alternative will be used for computing confidence intervals. Here, the standard errors address replicated random samples from a population for which the randomized population is arguably a random sample. *P*-values under the alternative might be useful since they will agree with the confidence interval, at least when both are based on normal approximations. Throughout this chapter, we take the approach that *p*-values will be computed under the null, with corresponding confidence intervals computed under the alternative hypothesis. The user can specify  $\alpha$  (**ALPHA**) to determine the confidence-level of the  $100(1 - \alpha)\%$  confidence intervals. By default, **ALPHA = 0.05**.

A single numeric variable can be provided to indicate strata (**STRATA**), and there are options to specify when covariate adjustment occurs relative to accounting for strata (**COMBINE**). By default, **COMBINE = NONE**, which assumes no stratification. When strata are present, the recommendation for when to account for covariate adjustment is typically determined by sample size. See Appendix A.6 for guidance on appropriate choices for **COMBINE**. How the estimates from the strata are combined is determined by **C**, which is the exponent for a Mantel-Haenszel weight. When **C = 1** (default), Mantel-Haenszel weights are used to combine strata. This considers the sample size of each stratum when combining estimates. When **C = 0**, strata are weighted equally, effectively ignoring the size of each stratum. Values of **C** between 0 or 1 put lesser or greater emphasis on strata sample sizes, respectively.

For computing resampling-based exact analyses (**EXACT**, **NO** by default), the user can specify a seed (**SEED**, 0 by default) to generate reproducible analyses, the number of simulations to conduct (**NREPS**), and the amount of memory to allocate to the symbol space (**SYMSIZE**, 20000 by default). Whether macro details are written to the SAS log or not is controlled by the **DETAILS** option (**YES** by default). For lengthy simulations, the user should consider setting **DETAILS = NO**; specifying **OPTIONS**

`NONOTES`; and using a SAS macro variable to report simulation progress. See Program 2.18 for an example.

The input data set name (`DSNIN`) specifies a single data set containing analysis data. The prefix for output data sets (`DSNOUT`) enables the user to customize the names of the output data sets generated by `%NParCov4`; see Appendix B.1.15 for a complete list of output data sets.

Appendix B.1 provides details for all `%NParCov4` macro options, and Appendix B.2 describes the contents of the generated output data sets.

## 2.4 Analysis of ordinal endpoints using a linear model

---

This section and the next five sections provide illustrations of the nonparametric, randomization-based methodology by applying it to the four clinical trial examples defined earlier in the chapter.

This section applies the randomization-based methodology to the analysis of an ordinal endpoint in Case study 1. Suppose that the trial's sponsor wants to compare the average response for the ordinal respiratory outcome at Visit 1 between the active and placebo treatments with adjustment for gender, age, and baseline score. We can perform a stratified analysis using centers as strata, weighting each center using Mantel-Haenszel weights ( $C = 1$ ). Further, we combine treatment estimates across strata prior to covariance adjustment using `COMBINE = FIRST`, assuming that the sample size per stratum is small. The `%NParCov4` call for this example is provided in Program 2.2.

### PROGRAM 2.2 Linear Model with Adjustment for Covariates for Visit 1

```
%NParCov4(OUTCOMES = V1, COVARS = gender age baseline, C=1,
HYPOTH = NULL, STRATA = center, TRTGRPS = treatment, TRANSFORM = NONE,
COMBINE = FIRST, EXACT = YES, NREPS = 5000, SEED = 36, DSNIN = RESP,
DSNOUT = OUTDAT);
```

In Program 2.2, `HYPOTH = NULL` so that we can first obtain  $p$ -values of interest for our analysis. The option `DSNOUT = OUTDAT` sets the value `OUTDAT` as the prefix for the output data sets. Throughout this chapter, we assume `DSNOUT = OUTDAT` and `C = 1`. We refrain from repeating these options in the programs below unless the values change. For this particular example, the following data sets are created:

1. `_OUTDAT_BETASAMP`, which contains the observed and `NREPS = 5000` permutation-based estimates of the treatment effect.
2. `_OUTDAT_COVBETA`, which contains the covariance matrix of the covariate-adjusted treatment effects for the outcomes.
3. `_OUTDAT_COVTEST`, which contains the criterion for covariate imbalance.
4. `_OUTDAT_DEPTEST`, which contains the treatment estimate, its standard error, and a two-sided asymptotic  $p$ -value.
5. `_OUTDAT_EXACT`, which contains one- and two-sided  $p$ -values for an essentially-exact analysis of `NREPS = 5000` permutation samples using `SEED = 36`.

The data set `_OUTDAT_COVTEST` contains the criterion for covariate imbalance  $Q=6.46$ , which, with sufficient sample size, is distributed  $\chi^2_{(3)}$ . The  $p$ -value for the criterion is 0.0911. While the exact  $p$ -value based on 5000 permutation samples is 0.0920, which is not contrary to random imbalance for the distributions of covariates

between the two treatments. In other words, randomization provided statistically similar distributions of covariates across the strata.

This criterion should not be considered a test for validity of nonparametric, randomization-based analysis of covariance, but merely as a measure of the degree of imbalance among covariates between the treatments. Due to the randomization requirement of the methodology, significant imbalance detected among the covariates should be interpreted as a chance finding. Adjusting for the covariates in the analysis removes this imbalance; the results of an unadjusted analysis would be confounded with the imbalance among the covariates. Small *p*-values for the covariate imbalance criterion are more appropriately referred to as ‘atypical relative to “chance” rather than “significant.”’

Data set *\_OUTDAT\_DEPTEST* summarizes results for the treatment difference that is provided in Table 2.5. Outcome, Treatment Effect, Standard Error, Test Statistic, and *P*-value are represented by the *\_OUTDAT\_DEPTEST* variables OUTCOMES, BETA, SEBETA, Q\_J, and PVALUE.

**TABLE 2.5** Linear Model with Adjustment for Covariates for Visit 1

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
Visit 1	0.4008	0.1714	5.4690	0.0194

An exact *p*-value based on 5000 permutation data sets is 0.0162. If there was interest in confidence intervals, we could specify HYPOTH = ALT in Program 2.2 above. An additional data set, *\_OUTDAT\_CI*, is created. It contains asymptotic 95% confidence intervals. Exact 95% bias-corrected and accelerated ( $BC_a$ ) and percentile intervals are included in the *\_OUTDAT\_EXACT* data set. The asymptotic 95% confidence interval for the treatment estimate is (0.1001, 0.7531). The corresponding bootstrap  $BC_a$  interval is (0.0974, 0.7749), and the percentile interval is (0.0901, 0.7646) based on 5000 bootstrap data sets.

The treatment effect in Table 2.5 represents the increase in the overall score for visit 1 for the active treatment compared to placebo. Based on our above analysis for Visit 1, we see that there is a significant treatment effect when stratifying by center and adjusting for gender, age, and baseline score. The treatment estimate is positive, indicating that the active drug had a greater positive effect, by about 0.40 on average, than placebo. The treatment estimate of approximately 0.40 could also be interpreted as corresponding to a  $0.40/4 = 0.10$  difference between treatments for the average of the cumulative probabilities of a score of at least 1, at least 2, at least 3, and at least 4. In contrast, the model without adjustment for covariates (Program 2.3) provides the results in Table 2.6.

### PROGRAM 2.3 Linear Model without Adjustment for Covariates for Visit 1

```
%NParCov4(OUTCOMES = V1, COVARS = , HYPOTH = NULL, STRATA = center,
TRTGRPS = treatment, TRANSFORM = NONE, COMBINE = FIRST, EXACT = YES,
NREPS = 5000, SEED = 36, DSNIN = RESP);
```

**TABLE 2.6** Linear Model without Adjustment for Covariates for Visit 1

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
Visit 1	0.3935	0.2032	3.7497	0.0528

An exact *p*-value based on 5000 permutation data sets is 0.0542. For confidence intervals, we could specify HYPOTH = ALT in Program 2.3 above. The asymptotic 95% confidence interval for the treatment estimate is (0.0024, 0.7846). Here, the  $BC_a$  and percentile intervals are (0.0197, 0.7906) and (0.0131, 0.7799), respectively, based on 5000 bootstrap data sets.

Note that for this particular example, the interpretation of the *p*-value and the confidence intervals don't coincide. This phenomenon does occur from time to time. This is because the *p*-value is based on re-randomizations of the population, while the confidence interval is based on repeated random sampling. This phenomenon can also occur for  $2 \times 2$  tables where the *p*-value for Fisher's exact test exceeds 0.05, but a confidence interval for the risk difference excludes 0. In this setting, Fisher's exact test is a randomization method based on re-randomizations of the population. But the confidence interval for the risk difference is based on repeated random sampling from a binomial distribution. The good news is that the result for the *p*-value and the confidence interval agree most of the time in terms of indicating whether a treatment effect exists. Otherwise, when the *p*-values and confidence intervals provide conflicting information, the *p*-value from the randomization test is based on more realistic and minimal assumptions, and should take precedence.

It is clear from the results of Tables 2.5 and 2.6 that adjustment for covariates provides a more sensitive comparison between the two treatments. To further illustrate the importance of covariate adjustment and the power of the nonparametric, randomization-based analysis of covariance methodology, Table 2.7 presents asymptotic 95% confidence intervals and interval widths for unstratified analyses similar to those presented in Programs 2.2 and 2.3, alongside comparable parametric analyses generated using PROC MIXED (Program 2.4). Intervals for both methods are non-significant or significant for the unadjusted or adjusted analysis, respectively. Further, in both the adjusted and unadjusted case, nonparametric, randomization-based analysis of covariance provides a narrower confidence interval than the model-based analysis, with greater precision observed when adjusting for covariates.

**TABLE 2.7 Estimated Treatment Effects and Asymptotic 95% Confidence Intervals by Method**

Method	Estimate	Lower Limit	Upper Limit	Width
Regression: Unadjusted	0.3996	-0.0129	0.8122	0.8251
NParCov: Unadjusted	0.3935	0.0024	0.7846	0.7822
Regression: Adjusted	0.4456	0.0995	0.7918	0.6923
NParCov: Adjusted	0.4266	0.1001	0.7531	0.6530

#### **PROGRAM 2.4 Parametric Analyses with and without Adjustment for Covariates for Visit 1**

```

proc mixed data = resp;
  ods output estimates = mixeddc;
  model v1 = gender age baseline treatment;
  estimate "Adjusted Treatment Effect" treatment 1;
run;

proc mixed data = resp;
  ods output estimates = mixed;
  model v1 = treatment;
  estimate "Unadjusted Treatment Effect" treatment 1;
run;

```

Suppose there is interest in examining the treatment effect across all 4 visits while adjusting for covariates and accounting for strata (Program 2.5).

#### **PROGRAM 2.5 Linear Model with Adjustment for Covariates for All Visits**

```
%NParCov4(OUTCOMES = V1 V2 V3 V4, HYPOTH = NULL,
COVARS = gender age baseline, STRATA = center, TRANSFORM = NONE,
TRTGRPS = treatment, COMBINE = FIRST, DSNIN = RESP);
```

The output is provided in Table 2.8.

**TABLE 2.8** Linear Model with Adjustment for Covariates for All Visits

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
Visit 1	0.4008	0.1714	5.4690	0.0194
Visit 2	0.9516	0.2213	18.4901	< 0.0001
Visit 3	0.8160	0.2386	11.6948	0.0006
Visit 4	0.6175	0.2377	6.7513	0.0094

The treatment estimates from `_OUTDAT_DEPTEST` and the covariance matrix from the data set `_OUTDAT_COVBETA` can be used to test more complex hypotheses. We could test the global hypothesis of a non-null treatment effect at the 4 visits,

$$H_0 : f_1 = 0, f_2 = 0, f_3 = 0, f_4 = 0,$$

where  $f_i$  is the treatment effect at the  $i$ th visit. The test statistic

$$Q_{Global} = \hat{\beta}^\top C^\top (C V_{\hat{\beta}} C^\top)^{-1} C \hat{\beta}$$

is distributed  $\chi^2_{(4)}$ , where  $C = I_4$  (an identity matrix of order four), and  $V_{\hat{\beta}}$  is the covariance matrix of the covariate-adjusted treatment effects of visits 1 to 4 for the overall response. The SAS/IML code in Program 2.6 performs the analysis.

## PROGRAM 2.6 Global Test for Treatment Response Across Visits

```
proc iml;
  use _OUTDAT_DEPTEST;
    read all var{beta} into B;
  close;
  use _OUTDAT_COVBETA;
    read all var{V1 V2 V3 V4} into V;
  close;
  C = I(4);
  Q = t(B) * t(C) * inv(C * V * t(C)) * C * B;
  p = 1 - probchi(Q, nrow(C));
  print Q p;
quit;
```

With  $Q_{Global} = 19.44$  and  $p = 0.0006$ , the analysis suggests that a non-null treatment effect exists for at least one visit. Note that the 4 degree-of-freedom test  $Q_{Global}$  would tend to have much less power than the 1 degree-of-freedom test that uses the average of the treatment differences across the visits; this would also be considered a global test of a treatment effect across the 4 visits.

An additional analysis can test the null hypothesis that the treatment effect does not differ across study visits,

$$H_0 : f_1 = f_2 = f_3 = f_4.$$

Here, the result of the test statistic and  $p$ -value for treatment  $\times$  visit should be based on the variance under the alternative rather than the variance under the null, since the treatment effects are already understood to be non-null. In this case, the test statistic

$$Q_{Visit} = \hat{\beta}^\top C^\top (C V_{\hat{\beta}} C^\top)^{-1} C \hat{\beta}$$

is distributed  $\chi^2_{(3)}$ , where

$$C = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

Changing HYPOTH = ALT in Program 2.5 and setting

```
C = {1 0 0 -1, 0 1 0 -1, 0 0 1 -1};
```

in Program 2.6 provides  $Q_{Treatment \times Visit} = 12.57$  and  $p = 0.0057$ . This suggests that the treatment effect differs across visits. In Section 2.16, we demonstrate how to generate essentially exact results for analyses with more than one response.

## 2.5 Analysis of binary endpoints

---

Case study 1 will be used to illustrate the analysis of binary endpoints with appropriate covariate adjustment. Specifically, in this section, we will analyze the binary response of (good, excellent) versus (terrible, poor, fair) at Visit 1. If the primary interest is in the test of the treatment effect, the most appropriate covariance matrix is the one that applies to the randomization distribution under the null hypothesis of no treatment difference. An exact  $p$ -value can be obtained; under the null hypothesis, all resampled data sets have the same covariance matrix. For illustration, we obtain weighted estimates across strata prior to covariance adjustment. The %NParCov4 call for this example is provided in Program 2.7.

### PROGRAM 2.7 Difference in Proportions with Adjustment for Covariates for Visit 1

```
data RESP;
  set RESP;
  v1goodex = (V1 in (3,4));
run;

%NParCov4(OUTCOMES = v1goodex, COVARS = gender age baseline,
HYPOTH = NULL, STRATA = center, TRTGRPS = treatment, TRANSFORM = NONE,
COMBINE = FIRST, EXACT = YES, NREPS = 5000, SEED = 78, DSNIN = RESP);
```

The criterion for covariate imbalance  $Q = 6.46$  is distributed  $\chi^2_{(3)}$ . The asymptotic  $p$ -value for the criterion is 0.0911, and the exact  $p$ -value provided in \_OUTDAT\_COVTEST based on 5000 permutation samples is 0.0922. It appears that randomization provided reasonably equal distributions of covariates across the strata. With TRANSFORM = NONE, the model provides the treatment estimate for the difference in proportions in Table 2.9.

**TABLE 2.9** Linear Model with Adjustment for Covariates for Visit 1

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
Visit 1	0.1839	0.0781	5.5455	0.0185

The exact  $p$ -value for the treatment effect provided in \_OUTDAT\_EXACT, based on 5000 permutation samples, is 0.0164. Therefore, active treatment provides a higher proportion of good or excellent responses when adjusting for gender, age, and baseline score. If there was interest in confidence intervals for the odds ratio, we could specify HYPOTH = ALT and TRANSFORM = LOGISTIC in the macro call above. However, this analysis causes an error for %NParCov4! What went wrong? The outcome was rare enough that some of the resampled data sets had no successes for at least one treatment arm within strata, and this caused the estimation to fail. If the outcome is expected to be rare, caution should be taken in specifying the analysis. For example, the following analysis adjusting for center as a covariate generates exact results (Program 2.8).

**PROGRAM 2.8 Logistic Model with the Addition of Center as a Covariate for Visit 1**

```
%NParCov4(OUTCOMES = v1goodex, COVARS = center gender age baseline,
HYPOTH = ALT, STRATA = NONE, TRTGRPS = treatment, TRANSFORM = LOGISTIC,
EXACT = YES, NREPS = 5000, SEED = 78, DSNIN = RESP);
```

The data set \_OUTDAT\_CI is created for all models and contains asymptotic confidence intervals for the treatment parameters when HYPOTH = ALT. However, an additional data set \_OUTDAT\_RATIOCI is produced here. It contains the asymptotic confidence intervals for the odds ratios. The odds ratio and asymptotic 95% confidence interval obtained from the data set \_OUTDAT\_RATIOCI is 2.2707 (1.2086, 4.2665). The BC<sub>a</sub> and percentile intervals for the odds ratio are (1.1510, 4.6950) and (1.1681, 4.7672), respectively, based on 5000 samples.

Ideally, the first analysis should be regenerated using the new options (center as a covariate) to be consistent with the analysis from Program 2.8. The appropriate code is presented in Program 2.9 (results not shown).

**PROGRAM 2.9 Difference in Proportions with Center as a Covariate for Visit 1**

```
%NParCov4(OUTCOMES = v1goodex, COVARS = center gender age baseline,
HYPOTH = ALT, STRATA = NONE, TRTGRPS = treatment, TRANSFORM = NONE,
EXACT = YES, NREPS = 5000, SEED = 78, DSNIN = RESP);
```

## 2.6 Analysis of ordinal endpoints using a proportional odds model

---

Case study 1 can also be used to illustrate an analysis strategy based on a proportional odds model. Rather than dichotomize the ordinal response in this clinical trial, we can analyze the outcome using a proportional odds model. There are few responses for poor and terrible so these categories will be combined for analysis. We can define 3 binary indicators to indicate excellent response or not, at least good response or not, or at least fair response or not (Program 2.10).

**PROGRAM 2.10 Proportional Odds Model with Adjustment for Covariates for Visit 1**

```
data RESP;
set RESP;
v1ex = (V1 in (4));
v1goodex = (V1 in (3,4));
v1fairgoodex = (V1 in (2,3,4));
run;

%NParCov4(OUTCOMES = v1ex v1goodex v1fairgoodex, COVARS = gender age baseline,
HYPOTH = NULL, STRATA = center, TRTGRPS = treatment, TRANSFORM = PODDS,
COMBINE = FIRST, DSNIN = RESP);
```

A model where TRANSFORM = PODDS generates an additional data set \_OUTDAT\_HOMOGEN that contains a test for the appropriateness of the proportional odds assumption. The test of proportional odds is  $Q = 3.69$ , which is distributed  $\chi^2_{(2)}$  and provides a  $p$ -value of 0.1578. Thus, we do not reject the assumption of proportional odds. Further, the covariate imbalance criterion, which in this case is a joint test of proportional odds and covariate imbalance (for gender, age, and baseline), is borderline ( $p=0.0709$ ). The model provides the following treatment estimate of the log odds ratio comparing active treatment to placebo in Table 2.10.

**TABLE 2.10 Proportional Odds Model with Adjustment for Covariates for Visit 1**

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
Visit 1	0.6233	0.3046	4.1857	0.0408

When HYPOTH = ALT is changed in Program 2.10, the data set \_OUTDAT\_RATIOCI contains odds ratio and asymptotic 95% confidence interval of 1.9548 (1.0455, 3.6548). Therefore, the odds of a positive response in the test treatment are greater than for placebo. Finally, had the test of homogeneity been statistically significant, the proportional odds assumption would not have applied for this analysis. In this case, the full model could be obtained by using TRANSFORM = LOGISTIC in the macro call above to estimate separate odds ratios for each binary outcome. See Program 2.11 for a sample program (results not shown).

**PROGRAM 2.11 Model with Adjustment for Covariates for Visit 1 - Proportional Odds Does Not Apply**

```
%NParCov4(OUTCOMES = v1ex v1goodex v1fairgoodex, COVARS = gender age baseline,
HYPOTH = ALT, STRATA = center, TRTGRPS = treatment, TRANSFORM = LOGISTIC,
COMBINE = FIRST, DSNIN = RESP);
```

## 2.7 Analysis of continuous endpoints using the log-ratio of two means

---

Analysis of continuous endpoints in the context of the nonparametric, randomization-based methodology can be illustrated using the osteoporosis trial from Case study 2. The analysis focuses on comparing the yearly risk of fractures between the two treatments using a risk ratio. We calculate the yearly risk for each subject as the total number of fractures divided by 3 years (or 2.5 years for the 7 subjects who discontinued early). Treatment (active = 2, placebo = 1) and center are recoded to be numeric. For illustration, we take weighted averages across the strata prior to taking the log transformation (based on the natural logarithm) and applying covariance adjustment (Program 2.12).

**PROGRAM 2.12 Log-Ratio of Two Means with Adjustment for Covariates**

```
%NParCov4(OUTCOMES = YEARLYRISK, COVARS = age, HYPOTH = NULL,
STRATA = center, TRTGRPS = treatment, TRANSFORM = LOGRATIO,
COMBINE = PRETRANSFORM, EXACT = YES, NREPS = 5000, SEED = 22,
DSNIN = FRACTURE);
```

The covariate imbalance criterion for this example is  $Q = 0.0536$  that is distributed  $\chi^2_{(1)}$  with  $p = 0.8170$ . The exact  $p$ -value is 0.8286. It appears that randomization provided a reasonably equal distribution of age between treatment groups. The model generates the treatment estimate of the log-ratio of the two means. It is presented in Table 2.11.

**TABLE 2.11 Log-Ratio of Two Means with Adjustment for Covariates**

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
Yearly Risk	-0.5803	0.3006	3.7259	0.0536

The exact  $p$ -value is 0.0626 based upon 5000 samples. If there was interest in confidence intervals for the log-ratio of the means, we could specify HYPOTH = ALT in the macro call above. Exponentiating the treatment estimate provides the

ratio 0.5600 which has 95% confidence interval (0.3068, 1.0222). Despite fractures being  $(1/0.56 - 1) \times 100\% = 79\%$  more likely per year for subjects on placebo compared to active drug, this difference is of borderline significance when adjusting for age. The  $BC_a$  and percentile intervals for the risk ratio are (0.2823, 0.9999) and (0.2882, 1.0209), respectively, based on 5000 samples. Here, the  $BC_a$  intervals imply a significant difference between placebo and active drug that conflicts with the asymptotic and percentile intervals. Doubling the number of bootstrap samples provides a  $BC_a$  interval that includes 1. Note that had there been greater variability in the duration of exposure for each subject, we could have analyzed with the methods in the following section.

## 2.8 Analysis of count endpoints using log-incidence density ratios

---

Several of the 40 clinical sites in the aneurysmal subarachnoid hemorrhage trial from Case study 3 had low enrollment, with fewer than two patients per treatment in some cases. In order to analyze under the alternative hypothesis, we will ignore center as a stratification variable. For each treatment, the incidence density will be the number of treatment-emergent, serious adverse events that a subject experienced, normalized by the number of days that he or she was in the trial. The ratio of the incidence densities (nicardipine over placebo) will be treated as the primary endpoint in the trial. Treatment estimates are adjusted for age and gender in the macro call of Program 2.13.

### PROGRAM 2.13 Log-Incidence Density Ratios with Adjustment for Covariates

```
%NParCov4(OUTCOMES = count, COVARS = age male, EXPOSURES = days,
HYPOTH = NULL, TRTGRPS = treatment, TRANSFORM = INCDENS,
EXACT = YES, NREPS = 5000, SEED = 77, DSNIN = nic);
```

The covariate imbalance criterion for this example is  $Q = 0.81$ , which is distributed  $\chi^2_{(2)}$  with the  $p$ -value of 0.666. The exact  $p$ -value is 0.6810. It appears that randomization provided a reasonably equal distribution of age and gender between treatment groups. The treatment estimate of the log-incidence density ratio is presented in Table 2.12.

**TABLE 2.12** Log-Incidence Density Ratios with Adjustment for Covariates

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
Incidence Density	-0.0718	0.1428	0.2526	0.6153

The exact  $p$ -value is 0.6104 based on 5000 samples. When **HYPOTH = ALT** is changed in Program 2.13, the data set **\_OUTDAT\_RATIOCI** contains incidence density ratio 0.9307 which has 95% confidence interval (0.7007, 1.2363). Treatment-emergent, serious adverse events are  $(1/0.9307 - 1) \times 100\% = 7.4\%$  more likely per day for subjects on placebo compared to nicardipine. This difference is not statistically significant when adjusting for the covariates. The  $BC_a$  and percentile intervals for the incidence density ratio are (0.6973, 1.2228) and (0.7009, 1.2303), respectively, based on 5000 samples.

## 2.9 Analysis of time-to-event endpoints

---

This section focuses on the colorectal cancer trial from Case study 4. First, we conduct the analysis of the primary time-to-event endpoint in the trial (time to death) using log-rank scores and adjusting for the number of days from initial cancer diagnosis to start of study drug administration and the sum of the longest lesion diameters in millimeters at baseline. We perform a stratified analysis of overall survival based upon the KRAS gene, weighting each center using Mantel-Haenszel weights. Further, we combine treatment estimates across strata prior to covariance adjustment (Program 2.14). Note that for time-to-event outcomes, the time until the event occurs or when the patient is censored (os\_time) is provided in EXPOSURES. The binary outcome of whether a patient experienced an event or not (here, os\_event, where 1 implies an event occurred, and 0 implies the patient was censored) is provided in OUTCOMES.

### **PROGRAM 2.14    Overall Survival Using Log-rank Scores with Adjustment for Covariates**

```
%NParCov4(OUTCOMES = os_event, COVARS = days length, EXPOSURES = os_time,
STRATA = mutant, HYPOTH = NULL, TRTGRPS = treatment, TRANSFORM =
LOGRANK, COMBINE = FIRST, EXACT = YES, NREPS = 5000, SEED = 45,
DSNIN = PROSTATE, DSNOUT = OS);
```

The criterion for covariate imbalance  $Q = 1.19$  is distributed  $\chi^2_{(2)}$ . The asymptotic  $p$ -value for the criterion is 0.5514, and the exact  $p$ -value based on 5000 permutation samples is 0.5484. It appears that randomization provided reasonably equal distributions of covariates between the two treatments. The treatment estimate, which represents the difference in average log-rank scores for overall survival, is presented in Table 2.13.

**TABLE 2.13    Overall Survival Using Log-rank Scores with Adjustment for Covariates**

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
Overall Survival	0.0299	0.1749	0.0292	0.8642

The exact  $p$ -value based on 5000 permutation samples is 0.8712. The positive treatment estimate implies that the treatment has decreased overall survival from colorectal cancer, relative to control. However, this result is not significant so we conclude that there is no difference between the treatments when adjusting for the covariates. A similar analysis for the other time-to-event endpoint in the colorectal cancer trial (time to disease progression) is performed in Program 2.15.

### **PROGRAM 2.15    Progression-Free Survival Using Log-rank Scores with Adjustment for Covariates**

```
%NParCov4(OUTCOMES = pfs_event, COVARS = days length, EXPOSURES = pfs_time,
STRATA = mutant, HYPOTH = NULL, TRTGRPS = treatment, TRANSFORM =
LOGRANK, COMBINE = FIRST, EXACT = YES, NREPS = 5000, SEED = 45,
DSNIN = PROSTATE, DSNOUT = PFS);
```

In this case, the criterion for covariate imbalance  $Q = 1.19$ , which is distributed  $\chi^2_{(2)}$ . The asymptotic  $p$ -value for the criterion is 0.5514, and the exact  $p$ -value based on 5000 permutation samples is 0.562. It appears that randomization provided reasonably equal distributions of covariates between the two treatments. The treatment estimate, which represents the difference in average log-rank scores for progression-free survival, is presented in Table 2.14.

**TABLE 2.14 Progression-Free Survival Using Log-rank Scores with Adjustment for Covariates**

Outcome	Treatment Effect	Standard Error	Test Statistic	P-value
PFS	0.0090	0.1750	0.0026	0.9592

The exact  $p$ -value based on 5000 permutation samples is 0.9600. The positive treatment estimate implies that the treatment has decreased progression-free survival from colorectal cancer, relative to control. However, this result is not significant, so we conclude that there is no difference between the treatments when adjusting for the covariates.

A joint test of the overall and progression-free survival can be performed with multiple calls to `%NParCov4` to test the hypothesis

$$H_0 : f_{os} = 0, f_{pfs} = 0,$$

where  $f_{os}$  and  $f_{pfs}$  are the treatment effects for overall and progression-free survival, respectively. The `_OUTDAT_SURV` data sets from the two runs above provide the log-rank scores (`LOGRANK_OS_EVENT` or `LOGRANK_PFS_EVENT`). These data sets can be merged together, and a joint analysis of overall and progression-free survival can be performed using Program 2.16.

**PROGRAM 2.16 Joint Analysis of Overall and Progression-Free Survival Using Log-rank Scores with Adjustment for Covariates**

```
%NParCov4(OUTCOMES = logrank_os_event logrank_pfs_event, COVARS = days length,
EXPOSURES = , STRATA = mutant, HYPOTH = NULL, TRTGRPS = treatment,
TRANSFORM = NONE, COMBINE = FIRST, EXACT = YES, NREPS = 5000, SEED = 45,
DSNIN = PROSTATE);
```

Notice that `EXPOSURES` is left unspecified since these values have already been accounted for in the `LOGRANK_OS_EVENT` and `LOGRANK_PFS_EVENT` outcomes. Further, no transformation is needed, so `TRANSFORM = NONE`. Defining  $C = I_2$ , an identity matrix of order two, and using similar methods and code as those for Program 2.6, the two degree-of-freedom bivariate test is computed in Program 2.17.

**PROGRAM 2.17 Bivariate Test of Overall and Progression-Free Survival**

```
proc iml;
  use _OUTDAT_DEPTEST;
    read all var{beta} into B;
  close;
  use _OUTDAT_COVBETA;
    read all var{logrank_os_event logrank_pfs_event} into V;
  close;
  C = I(2);
  Q = t(B) * t(C) * inv(C * V * t(C)) * C * B;
  p = 1 - probchi(Q, nrow(C));
  print Q p;
quit;
```

Here,  $Q_{Bivariate} = 0.0295$  with asymptotic  $p = 0.9854$ . It is natural to ask if an exact  $p$ -value is available for this example. As described in Appendix B, the exact methodologies of `%NParCov4` are limited to a single outcome variable. However, it is possible to write additional code to repeatedly call `%NParCov4`. Though repeated calls to `%NParCov4` might not be the most efficient procedure programmatically, it does enable the analyst to compute exact results for more general problems.

In Program 2.18, the  $Q_{Bivariate}$  statistic is computed for each permutation sample, and the proportion of these  $Q_{Bivariate}$  statistics that is greater than or equal to the  $Q_{Bivariate}$  statistic computed from the observed sample constitutes the  $p$ -value.

### **PROGRAM 2.18    Exact Bivariate Test of Overall and Progression-Free Survival**

```
%let nreps = 5000;
%let seed = 50;

options nonotes;
proc sort data = both;
   by treatment patient;
run;

proc multtest permutation nocenter noprint seed=&seed nsample=&nreps
   outsamp=samp_null1(drop=_obs_) data=both;
   class treatment;
   strata mutant;
   test mean(logrank_os_event logrank_pfs_event days length);
run;

data samp_null2;
   set samp_null1;
   *** convert treatment back to numeric ***;
   treatment =_class_+0;
   *** convert strata back to numeric ***;
   mutant =_stratum_+0;
   drop _class_;
run;

*** sample 0 is the observed sample ***;
data samp_null3;
   set both(in=a) samp_null2;
   if a then _sample_=0;
run;

%macro BIVARIATE;
%do rep = 0 %to &nreps;
   data temp;
      set samp_null3;
      where _sample_ = &rep;
   run;

   %put PROGRESS = &rep;

   %NPARCOV4(outcomes = logrank_os_event logrank_pfs_event,
   covars = days length,
   trtgrps = treatment,
   strata = mutant,
   hypoth = NULL,
   transform = NONE,
   combine = FIRST,
   c = 1,
   dsnin = temp,
   dsnout = outdat,
   exact = NO,
   details = NO);

   %if &rep = 0 %then %do;
```

```

data allbetas;
    set _outdat_deptest(keep = outcomes beta);
    _sample_ = &rep;
run;

data allcovs;
    set _outdat_covbeta;
    _sample_ = &rep;
run;
%end;
%else %do;
    data allbetas;
        set allbetas _outdat_deptest(keep = outcomes beta in = inb);
        if inb then _sample_ = &rep;
    run;

    data allcovs;
        set allcovs _outdat_covbeta(in = inb);
        if inb then _sample_ = &rep;
    run;
%end;

proc datasets nolist;
    delete temp;
    quit;
%end;
%mend BIVARIATE;

%BIVARIATE;

proc iml;
    use allbetas;
    read all var{beta} into B;
    close;
    use allcovs;
    read all var{logrank_os_event logrank_pfs_event} into V;
    close;
    C = I(2);
    Q0 = t(B[1:2]) * C * inv(C * V[1:2,] * t(C)) * C * B[1:2];
    twosided = 0;
    do i = 2 to &nreps+1;
        Btemp = B[1+(i-1)*ncol(V):i*ncol(V)];
        Vtemp = V[1+(i-1)*ncol(V):i*ncol(V),];
        Q = t(Btemp) * t(C) * inv(C * Vtemp * t(C)) * C * Btemp;
        twosided = twosided + (Q >= Q0);
    end;
    twosided = twosided / &nreps;
    print Q0 twosided;
quit;

options notes;

```

Program 2.18 generates 5000 permutation samples to compute an essentially exact  $p$ -value = 0.9836. DETAILS = NO and OPTIONS NONOTES are used to minimize the amount of information saved to the SAS log while these simulations are running. Note that in practice, an exact  $p$ -value is typically not useful when the asymptotic  $p$ -value exceeds 0.15. Computing an exact  $p$ -value here was for illustration more than anything else.

This entire analysis could have been conducted using Wilcoxon scores. An alternative analysis that might be of interest is to conduct a joint analysis of an endpoint using both Wilcoxon and log-rank scores. For example, in the above description, merely replace the progression-free survival call with an additional analysis for overall survival using Wilcoxon scores (`TRANSFORM = WILCOXON`) to test the null hypothesis

$$H_0 : f_{\text{Wilcoxon}} = 0, f_{\text{logrank}} = 0,$$

where  $f_{\text{Wilcoxon}}$  and  $f_{\text{logrank}}$  are the treatment effects for overall survival. They are computed using the Wilcoxon or log-rank scores, respectively. What is the rationale for such an analysis? In some instances, it may not be clear which test might be most appropriate for a given data set. For example, the log-rank test assumes proportional hazards between the two treatments that might not hold in practice. Wilcoxon scores place greater emphasis on earlier event times (Collett, 2015). If the proportional hazards assumption holds, it might be unclear which test might be more appropriate. In oncology, the log-rank test is often more powerful to assess long-term survival. However, in analgesic studies where interest is in earlier pain relief, the Wilcoxon test might be preferred. See Tangen and Koch (1999b) for an illustration of this particular bivariate test.

## 2.10 Summary

---

This chapter presented the nonparametric, randomization-based analysis of covariance methodology and its application to the analysis of clinical endpoints. The methodology was implemented using the `%NParCov4` macro that contains functionality for many of the analysis methods described to date, with the ability to compute essentially-exact  $p$ -values and confidence intervals. It is recommended that advanced randomization-based methods be specified as the primary analysis because of their minimal assumptions and the availability of essentially-exact  $p$ -values. Model-based methods, such as a logistic regression model in the case of binary outcomes, can be used in a supportive fashion to assess the effects of covariates, potential interactions of treatment with covariates, and the generalizability of findings to populations not represented by the distributions of subjects in a clinical trial.

## Acknowledgements

---

The authors extend their sincerest thanks to Elaine Kearney, Todd Schwartz, Hengrui Sun, and Laura Zhou for their careful review of this chapter.

## References

---

- Collett D. (2015). *Modelling Survival Data in Medical Research, Third Edition*. Boca Raton, FL: CRC Press.
- Davison A.C., Hinkley D.V. (1997). *Bootstrap Methods and Their Applications*. Cambridge, UK: Cambridge University Press.
- Efron B., Tibshirani R.J. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman and Hall.

- Eisenhauer E.A., Therasse P., Bogaerts J., Schwartz L.H., Sargent D., Ford R., Dancey J., Arbuck S., Gwyther S., Mooney M., Rubinstein L., Shankar L., Dodd L., Kaplan R., Lacombe D., Verweij J. (2009). New response evaluation criteria in solid tumors: Revised RECIST guideline (version 1.1). *European Journal of Cancer* 45, 228-247.
- Haley E.C., Kassell N.F., Torner J.C. (1993). A randomized controlled trial of high-dose intravenous nicardipine in aneurysmal subarachnoid hemorrhage. *Journal of Neurosurgery* 78, 537-547.
- Jennett B., Bond M. (1975). Assessment of outcome after severe brain damage: A practical scale. *Lancet* 1, 480-484.
- Kalbfleisch J.D., Prentice R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Kawaguchi A., Koch G.G., Wang X. (2011). Stratified multivariate Mann-Whitney estimators for the comparison of two treatments with randomization based covariance adjustment. *Statistics in Biopharmaceutical Research* 3, 217-231.
- Kawaguchi A., Koch G.G. (2015). sanon: An R package for stratified analysis with nonparametric covariate adjustment. *Journal of Statistical Software* 67:9, 1-37. Available at <http://www.jstatsoft.org/v67/i09/>.
- Koch G.G., Carr G.J., Amara I.A., Stokes M.E., Uryniak T.J. (1990). Categorical data analysis. In: Berry D.A., ed. *Statistical Methodology in the Pharmaceutical Sciences*. New York: Marcel Dekker.
- Koch G.G., Gillings D.B. (1983). Inference, design-based vs. model-based. In: Kotz S., Johnson N.L., Read C.B., eds. *Encyclopedia of Statistical Sciences*. New York: Wiley.
- Koch G.G., Tangen C.M. (1999). Nonparametric analysis of covariance and its role in non-inferiority clinical trials. *Drug Information Journal* 33, 1145-1159.
- Koch G.G., Tangen C.M., Jung J.W., Amara I.A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine* 17, 1863-1892.
- LaVange L.M., Durham T.A., Koch G.G. (2005). Randomization-based nonparametric methods for the analysis of multicentre trials. *Statistical Methods in Medical Research* 14, 281-301.
- LaVange L.M., Koch G.G. (2008). Randomization-based nonparametric analysis of covariance (ANCOVA). In: D'Agostino R.B., Sullivan L., Massaro J., eds. *Wiley Encyclopedia of Clinical Trials, First Edition*. Hoboken, NJ: Wiley.
- Moodie P.F., Saville B.R., Koch G.G., Tangen C.M. (2011). Estimating covariate-adjusted log hazard ratios for multiple time intervals in clinical trials using nonparametric randomization based ANCOVA. *Statistics in Biopharmaceutical Research* 3, 232-241.
- Saville B.R., Herring A.H., Koch G.G. (2010). A robust method for comparing two treatments in a confirmatory clinical trial via multivariate time-to-event methods that jointly incorporate information from longitudinal and time-to-event data. *Statistics in Medicine* 29, 75-85.
- Saville B.R., Koch G.G. (2013). Estimating covariate-adjusted log hazard ratios in randomized clinical trials using Cox proportional hazards models and nonparametric randomization based analysis of covariance. *Journal of Biopharmaceutical Statistics* 23, 477-490.
- Saville B.R., LaVange L.M., Koch G.G. (2011). Estimating covariate-adjusted incidence density ratios for multiple time intervals in clinical trials using nonpara-

- metric randomization based ANCOVA. *Statistics in Biopharmaceutical Research* 3, 242-252.
- Stokes M.E., Davis C.S., Koch G.G. (2012). *Categorical Data Analysis Using The SAS System, Third Edition*. Cary, NC: SAS Institute Inc. Data available at <http://support.sas.com/rnd/app/stat/cat/edition3/samples/index.html>.
- Tangen C.M., Koch G.G. (1999a). Complementary nonparametric analysis of covariance for logistic regression in a randomized clinical trial setting. *Journal of Biopharmaceutical Statistics* 9, 45-66.
- Tangen C.M., Koch G.G. (1999b). Nonparametric analysis of covariance for hypothesis testing with logrank and Wilcoxon scores and survival-rate estimation in a randomized clinical trial. *Journal of Biopharmaceutical Statistics* 9, 307-338.
- Tangen C.M., Koch G.G. (2000). Non-parametric covariance methods for incidence density analyses of time-to-event data from a randomized clinical trial and their complementary roles to proportional hazards regression. *Statistics in Medicine* 19, 1039-1058.
- Tangen C.M., Koch G.G. (2001). Non-parametric analysis of covariance for confirmatory randomized clinical trials to evaluate dose response relationships. *Statistics in Medicine* 20, 2585-2607.
- Zhao Y., Saville B.R., Zhou H., Koch G.G. (2016). Sensitivity analysis for missing outcomes in time-to-event data with covariate adjustment. *Journal of Biopharmaceutical Statistics* 26, 269-279.

## Appendix

---

### A Theoretical Details

#### A.1 Non-Parametric Randomization-Based Analysis of Covariance

Suppose there are two randomized treatment groups, and there is interest in a nonparametric, randomization-based comparison of  $r$  outcomes adjusting for  $t$  covariates within a single stratum. Let  $n_i$  be the sample size from a population of  $N_i >> n_i$  patients,  $\bar{y}_i$  be the  $r$ -vector of outcome means, and  $\bar{x}_i$  be the  $t$ -vector of covariate means of the  $i$ th treatment group,  $i = 1, 2$ . Let

$$f_i = \begin{bmatrix} g(\bar{y}_i) \\ \bar{x}_i \end{bmatrix},$$

where  $g(\cdot)$  is a twice-differentiable function. Further, define  $\text{cov}(f_i) = V_{f_i} = D_i V_i D_i$ , where

$$V_i = \text{cov} \left( \begin{bmatrix} \bar{y}_i \\ \bar{x}_i \end{bmatrix} \right)$$

and  $D_i$  is a diagonal matrix with the first derivatives  $g'(\bar{y}_i)$  corresponding to the first  $r$  cells of the diagonal with ones for the remaining  $t$  cells of the diagonal.

Define  $f = f_2 - f_1$  and  $V_f = V_{f_1} + V_{f_2}$  and let  $I_r$  be an identity matrix of dimension  $r$  and  $0_{(t \times r)}$  be a  $t \times r$  matrix of zeros and fit the model

$$\mathbb{E}[f] \hat{=} \begin{bmatrix} I_r \\ 0_{(t \times r)} \end{bmatrix} \hat{\beta} = X \hat{\beta}$$

using weighted least squares with weights based on  $V_f^{-1}$ , with  $\hat{=}$  meaning "is estimated by".  $\hat{\beta}$  is an  $r \times 1$  vector of covariate-adjusted treatment effects. The weighted least squares estimator is

$$\hat{\beta} = (X^\top V_f^{-1} X)^{-1} X^\top V_f^{-1} f$$

and a consistent estimator for  $\text{cov}(\hat{\beta})$  is  $(X^\top V_f^{-1} X)^{-1}$ . There are two ways to estimate  $V_f$ . Under the null hypothesis of no treatment difference,

$$V_i = \frac{1}{n_i(n_1 + n_2 - 1)} \left\{ \sum_{i=1}^2 \sum_{l=1}^{n_i} \begin{bmatrix} (y_{il} - \bar{y})(y_{il} - \bar{y})^\top & (y_{il} - \bar{y})(x_{il} - \bar{x})^\top \\ (x_{il} - \bar{x})(y_{il} - \bar{y})^\top & (x_{il} - \bar{x})(x_{il} - \bar{x})^\top \end{bmatrix} \right\},$$

where  $\bar{y}$  and  $\bar{x}$  are the means for the outcomes and covariates across both treatment groups, respectively. Under the alternative,

$$V_i = \frac{1}{n_i(n_i - 1)} \left\{ \sum_{l=1}^{n_i} \begin{bmatrix} (y_{il} - \bar{y}_i)(y_{il} - \bar{y}_i)^\top & (y_{il} - \bar{y}_i)(x_{il} - \bar{x}_i)^\top \\ (x_{il} - \bar{x}_i)(y_{il} - \bar{y}_i)^\top & (x_{il} - \bar{x}_i)(x_{il} - \bar{x}_i)^\top \end{bmatrix} \right\}.$$

In general, the variance under the null would be the principal structure for producing  $p$ -values for hypothesis testing under the null when the null corresponded to no difference between treatments (or more formally, the strong null that each patient would have the same outcome, regardless of the assigned treatment). The variance under the alternative will be used for computing confidence intervals. Here, the standard errors address replicated random samples from a population for which the randomized population is arguably a random sample.  $P$ -values under the alternative might be useful since they will agree with the confidence interval, at least when both are based on normal approximations. Throughout this chapter, we take the approach that  $p$ -values will be computed under the null, with corresponding confidence intervals computed under the alternative hypothesis.

A test to compare the two treatments for outcome  $y_j, j = 1, 2, \dots, r$ , with adjustment for covariates is

$$Q_j = \frac{\hat{\beta}_j^2}{V_{\hat{\beta}_j}}.$$

With sufficient sample size,  $Q_j$  is approximately distributed  $\chi_{(1)}^2$ . A criterion for the random imbalance of the covariates between the two treatment groups is

$$Q = (f - X\hat{\beta})^\top V_f^{-1} (f - X\hat{\beta})$$

which is approximately distributed  $\chi_{(t)}^2$ . This criterion should not be considered a test for validity of nonparametric, randomization-based analysis of covariance, but merely as a measure of the degree of imbalance among covariates between the treatments. Due to the randomization requirement of the methodology, significant imbalance detected among the covariates should be interpreted as a chance finding. Adjusting for the covariates in the analysis removes this imbalance. The results of an unadjusted analysis would be confounded with the imbalance among the covariates. Small  $p$ -values for the covariate imbalance criterion are more appropriately referred to as "atypical relative to chance" rather than "significant."

Under the alternative hypothesis, asymptotic confidence intervals for  $\hat{\beta}_j$  can be defined as  $\hat{\beta}_j \pm Z_{(1-\alpha/2)} V_{\hat{\beta}_j}^{1/2}$ , where  $Z_{(1-\alpha/2)}$  is a quantile from a standard normal random variable.

In many situations, a linear transformation of  $\bar{y}_i$  is appropriate, e.g.,  $g(\bar{y}_i) = \bar{y}_i$ . Here,  $D_i$  will be an identity matrix of dimension  $(r \times t)$ . In `%NParCov4`, the above methods correspond to `TRANSFORM = NONE` and `COMBINE = NONE`.

## A.2 Logistic Transformation

If  $\bar{y}_i$  consists of (0,1) binary outcomes, a logistic transformation of the elements of  $\bar{y}_i$  might be more appropriate. Let  $g(\bar{y}_i) = \text{logit}(\bar{y}_i)$  where  $\text{logit}(\cdot)$  transforms each

element  $\bar{y}_{ij}$  of  $\bar{y}_i$  to  $\log_e(\bar{y}_{ij}/(1 - \bar{y}_{ij}))$ . The first  $r$  elements of  $f$  correspond to

$$\log_e \frac{\bar{y}_{2j}(1 - \bar{y}_{1j})}{\bar{y}_{1j}(1 - \bar{y}_{2j})},$$

which is the  $\log_e$ -odds ratio of the  $j$ th outcome. Under the null hypothesis, the first  $r$  elements of the diagonal of  $D_i$  are  $\{\bar{y}_j(1 - \bar{y}_j)\}^{-1}$ , where  $\bar{y}_j$  is the mean for outcome  $y_j$  across both treatment groups. Under the alternative, the first  $r$  elements of the diagonal of  $D_i$  are  $\{\bar{y}_{ij}(1 - \bar{y}_{ij})\}^{-1}$ . The  $\hat{\beta}_j$  and associated confidence intervals can be exponentiated to obtain estimates and confidence intervals for the odds ratios. These methods correspond to **TRANSFORM = LOGISTIC** and **COMBINE = NONE**.

A single ordinal outcome with  $(r + 1)$  levels can be analyzed as  $r$  binary outcomes (cumulative across levels) using a proportional odds model. Logistic transformations are applied to the  $r$  outcomes as described above. A test of the homogeneity of the  $r$  odds ratios would be computed as

$$Q_c = \hat{\beta}^\top C^\top (CV_{\hat{\beta}}C^\top)^{-1} C \hat{\beta},$$

where  $C = [ I_{(r-1)} \quad -1_{(r-1)} ]$ ,  $I_{(r-1)}$  is an identity matrix of dimension  $(r - 1)$  and  $-1_{(r-1)}$  is a vector of length  $(r - 1)$  containing elements  $-1$ . With sufficient sample size,  $Q_c$  is approximately distributed  $\chi^2_{(r-1)}$ .

If homogeneity applies, you can use the simplified model

$$E[f] = \begin{bmatrix} 1_r \\ 0_t \end{bmatrix} \hat{\beta}_R = X_R \hat{\beta}_R.$$

For ordinal outcomes,  $\hat{\beta}_R$  would be the common odds ratio across cumulative outcomes in a proportional odds model. The criterion for random imbalance for this simplified model will be a joint test of random imbalance and proportional odds, which is approximately  $\chi^2_{(t+r-1)}$ . These methods correspond to **TRANSFORM = PODDS** and **COMBINE = NONE**. **%NParCov4** computes the test of homogeneity and the reduced model when **PODDS** is selected. Should the  $p$ -value for  $Q_c$  be such that it is  $\leq \alpha$ , the proportional odds assumption is violated, which implies that the full model is more appropriate (obtained with **TRANSFORM = LOGISTIC**).

### A.3 Log-Ratio of Two Outcome Vectors

Suppose there was interest in the analyses of the ratio of means while adjusting for the effect of covariates. Let  $g(\bar{y}_i) = \log_e(\bar{y}_i)$  where  $\log_e(\cdot)$  transforms each element  $\bar{y}_{ij}$  of  $\bar{y}_i$  to  $\log_e(\bar{y}_{ij})$ . The first  $r$  elements of  $f$  correspond to

$$\log_e(\bar{y}_{2j}) - \log_e(\bar{y}_{1j}) = \log_e \left( \frac{\bar{y}_{2j}}{\bar{y}_{1j}} \right),$$

which is the  $\log_e$ -ratio of the  $j$ th outcome. Under the null hypothesis, the first  $r$  elements of the diagonal of  $D_i$  are  $\bar{y}_j^{-1}$ , where  $\bar{y}_j$  is the mean for outcome  $y_j$  across both treatment groups. Under the alternative, the first  $r$  elements of the diagonal of  $D_i$  are  $\bar{y}_{ij}^{-1}$ . The  $\hat{\beta}_j$  and associated confidence intervals can be exponentiated to obtain estimates and confidence intervals for the ratios of means. These methods correspond to **TRANSFORM = LOGRATIO** and **COMBINE = NONE**.

### A.4 Log-Incidence Density Ratios

Suppose each of the  $r$  outcomes has an associated exposure so that  $\bar{e}_i$  is an  $r$ -vector of average exposures. Let

$$f_i = \begin{bmatrix} \log_e(\bar{y}_i) \\ \log_e(\bar{e}_i) \\ \bar{x}_i \end{bmatrix},$$

where  $\log_e(\cdot)$  transforms each element  $\bar{y}_{ij}$  of  $\bar{y}_i$  to  $\log_e(\bar{y}_{ij})$  and each element  $\bar{e}_{ij}$  of  $\bar{e}_i$  to  $\log_e(\bar{e}_{ij})$  and let

$$M_0 = \begin{bmatrix} I_r & -I_r & 0_{(r \times t)} \\ 0_{(t \times r)} & 0_{(t \times r)} & I_t \end{bmatrix},$$

where  $0_{(r \times t)}$  is an  $r \times t$  matrix of zeros and  $I_r$  and  $I_t$  are identity matrices of dimension  $r$  and  $t$ , respectively. Define

$$g_i = M_0 f_i = \begin{bmatrix} \log_e(\bar{y}_i) - \log_e(\bar{e}_i) \\ \bar{x}_i \end{bmatrix}$$

so that  $\text{cov}(g_i) = V_{g_i} = M_i V_i M_i^\top$ , where

$$V_i = \text{cov} \left( \begin{bmatrix} \bar{y}_i \\ \bar{e}_i \\ \bar{x}_i \end{bmatrix} \right)$$

and

$$M_i = \begin{bmatrix} D_{\bar{y}_i} & -D_{\bar{e}_i} & 0_{(r \times t)} \\ 0_{(t \times r)} & 0_{(t \times r)} & I_t \end{bmatrix}.$$

The first  $r$  elements of  $g_i$  are  $\log_e$ -incidence densities so that if  $g = g_2 - g_1$ , the first  $r$  elements of  $g$  correspond to

$$\log_e \left( \frac{\bar{y}_{2j}}{\bar{e}_{2j}} \right) - \log_e \left( \frac{\bar{y}_{1j}}{\bar{e}_{1j}} \right) = \log_e \left( \frac{\left( \frac{\bar{y}_{2j}}{\bar{e}_{2j}} \right)}{\left( \frac{\bar{y}_{1j}}{\bar{e}_{1j}} \right)} \right),$$

which is the  $\log_e$ -incidence density ratio of the  $j$ th outcome. Under the null hypothesis,  $D_{\bar{y}_i}$  is a diagonal matrix with diagonal elements  $\bar{y}_j^{-1}$ , and  $D_{\bar{e}_i}$  is a diagonal matrix with diagonal elements  $\bar{e}_j^{-1}$ , where  $\bar{y}_j$  and  $\bar{e}_j$  are the means for outcome  $y_j$  and exposure  $e_j$  across both treatment groups, respectively. Under the alternative hypothesis,  $D_{\bar{y}_i}$  is a diagonal matrix with diagonal elements  $\bar{y}_{ij}^{-1}$ . And  $D_{\bar{e}_i}$  is a diagonal matrix with diagonal elements  $\bar{e}_{ij}^{-1}$ . The  $\hat{\beta}_j$  and associated confidence intervals can be exponentiated to obtain estimates and confidence intervals for the incidence density ratios. These methods correspond to **TRANSFORM = INCdens** and **COMBINE = NONE**.

### A.5 Wilcoxon and Log-rank Scores for Time-to-Event Outcomes

Unlike Sections A.2, A.3, and A.4 where a transformation is applied to the means of one or more outcomes, **%NParCov4** can transform subject-level event flags and times into Wilcoxon or log-rank scores to perform analyses of time-to-event outcomes. Scores are defined from binary outcome flags (**OUTCOMES**) that define whether an event occurs (= 1) or not (= 0) and the exposure times (**EXPOSURES**) when the event occurs or the last available time at which an individual is censored (e.g., when the subject does not experience the event). The definitions are from the appendix of Tangen and Koch (1999b). The Wilcoxon scores are closely related to the Prentice generalized Wilcoxon scores presented in Kalbfleisch and Prentice (1980).

Suppose there are  $L$  unique failure times such that  $y_{(1)} < y_{(2)} < \dots < y_{(L)}$ . Assume that  $g_\zeta$  is the number of failures that occur at  $y_\zeta$  and that there are  $y_{\zeta_1}, y_{\zeta_2}, \dots, y_{\zeta_{\nu_\zeta}}$  censored values in  $[y_{(\zeta)}, y_{(\zeta+1)})$ , where  $\nu_\zeta$  is the number of censored subjects in the interval and  $\zeta = 1, 2, \dots, L$ . Further, assume  $y_{(0)} = 0, y_{(L+1)} = \infty$ , and  $g_0 = 0$ .

Log-rank scores are expressed as

$$c_\zeta = 1 - \sum_{\zeta'=1}^{\zeta} \frac{g_{\zeta'}}{N_{\zeta'}} \quad \text{and} \quad C_\zeta = - \sum_{\zeta'=1}^{\zeta} \frac{g_{\zeta'}}{N_{\zeta'}}$$

and  $C_0 = 0$ , where  $c_\zeta$  and  $C_\zeta$  are the scores for the uncensored and censored subjects, respectively, in the interval  $[y_{(\zeta)}, y_{(\zeta+1)})$  for  $\zeta = 1, 2, \dots, L$  where  $N_\zeta = n - \sum_{\zeta'=0}^{\zeta-1} (\nu_{\zeta'} + g_{\zeta'})$  is the number of subjects at risk at  $t_{(\zeta)} - 0$  and  $n = n_1 + n_2$  is the sample size. To conduct an unstratified analysis using log-rank scores, set **TRANSFORM** = **LOGRANK** and **COMBINE** = **NONE**.

Wilcoxon scores are expressed as

$$c_\zeta = 2 \prod_{\zeta'=1}^{\zeta} \left( \frac{N_{\zeta'} - g_{\zeta'}}{N_{\zeta'}} \right) - 1 \quad \text{and} \quad C_\zeta = \prod_{\zeta'=1}^{\zeta} \left( \frac{N_{\zeta'} - g_{\zeta'}}{N_{\zeta'}} \right) - 1$$

and  $C_0 = 0$ , where  $c_\zeta$  and  $C_\zeta$  are the scores for the uncensored and censored subjects, respectively, in the interval  $[y_{(\zeta)}, y_{(\zeta+1)})$  for  $\zeta = 1, 2, \dots, L$  where  $N_\zeta$  is the number of subjects at risk at  $t_{(\zeta)} - 0$ . To conduct an unstratified analysis using Wilcoxon scores, set **TRANSFORM** = **WILCOXON** and **COMBINE** = **NONE**. **%NParCov4** computes survival scores, summarizes as  $\bar{y}_i$  and analyzes as in Section A.1 using a linear transformation.

## A.6 Stratification

Stratification can be handled in one of three ways:

1. Estimate  $\hat{\beta}_h$  for each stratum  $h$ ,  $h = 1, 2, \dots, H$ , and combine across strata to obtain

$$\bar{\beta} = \frac{\sum_h w_h \hat{\beta}_h}{\sum_h w_h},$$

where

$$w_h = \left( \frac{n_{h1} n_{h2}}{n_{h1} + n_{h2}} \right)^c$$

and  $0 \leq c \leq 1$ , with  $c = 1$  for Mantel-Haenszel weights and  $c = 0$  for equal strata weights. A criterion for random imbalance is performed within each stratum and then summed to give an overall test of imbalance across all strata. This method is more appropriate for large sample size per stratum ( $n_{hi} \geq 30$  and preferably  $n_{hi} \geq 60$ ). This corresponds to **COMBINE** = **LAST**.

2. Form

$$f_w = \frac{\sum_h w_h f_h}{\sum_h w_h}$$

first and fit a model for covariance adjustment to determine  $\hat{\beta}_w$ . This method is more appropriate for small to moderate sample size per stratum. This corresponds to **COMBINE** = **FIRST**.

3. For transformations in Sections A.2, A.3, and A.4, define

$$\begin{bmatrix} \bar{y}_{w_i} \\ \bar{x}_{w_i} \end{bmatrix} = \left( \sum_h w_h \right)^{-1} \times \sum_h \left( w_h \begin{bmatrix} \bar{y}_{hi} \\ \bar{x}_{hi} \end{bmatrix} \right)$$

first, then proceed with the transformation and finally apply covariance adjustment. This method is more appropriate for small sample size per stratum and corresponds to **COMBINE** = **PRETRANSFORM**.

### A.7 Resampling-Based Exact Methods

%NParCov4 contains options for computing essentially exact  $p$ -values under HYPOTH = NULL and confidence intervals under HYPOTH = ALT for a single outcome. Under EXACT = YES and HYPOTH = NULL, %NParCov4 uses PROC MULTTEST with the PERMUTATION option to compute NREPS data sets. In this case, the sampling of records to either treatment (within strata) is performed without replacement. Under EXACT = YES and HYPOTH = ALT, %NParCov4 uses PROC MULTTEST with the BOOTSTRAP option and a BY statement for treatment to compute NREPS data sets where the sampling of records occurs with replacement by treatment (within strata). Further, under EXACT = YES and HYPOTH = ALT,  $n$  jackknife data sets are created where a single observation is deleted from each jackknife data set within strata, where  $n$  is the total number of observations in the analysis. The jackknife data sets are used in the computation of acceleration for the bias-corrected and accelerated ( $\text{BC}_a$ ) intervals.

Exact  $p$ -values are obtained for EXACT = YES and HYPOTH = NULL. An exact criterion for the random imbalance of covariates between the two treatment groups computes  $Q_m$ ,  $m = 1, 2, \dots, B$ , for each of  $B$  resampled data sets. The  $p$ -value is calculated as  $\#(Q_m \geq Q_0)/B$ , where  $Q_0$  is the criterion for random imbalance for the observed data set. Under the null hypothesis, the treatment parameter  $\hat{\beta}_m$ ,  $m = 1, 2, \dots, B$ , is computed for each resampled data set and is compared to the treatment parameter of the observed data set ( $\hat{\beta}_0$ ) to calculate  $p$ -values:

1. two-sided:  $\#(|\hat{\beta}_m| \geq |\hat{\beta}_0|)/B$ ,
2. one-sided lower:  $\#(\hat{\beta}_m \leq \hat{\beta}_0)/B$ ,
3. one-sided upper:  $\#(\hat{\beta}_m \geq \hat{\beta}_0)/B$ .

Under EXACT = YES and HYPOTH = ALT, bias-corrected and accelerated intervals ( $\text{BC}_a$ ) are computed using the  $B$  bootstrap samples and the  $n$  jackknife samples (Efron and Tibshirani, 1993). The  $100(1 - \alpha)\%$   $\text{BC}_a$  interval is the  $\alpha_1$  and  $\alpha_2$  percentiles of all  $B$  bootstrap estimates and is defined as  $\text{BC}_a = (\hat{\beta}_{(\alpha_1)}, \hat{\beta}_{(\alpha_2)})$ , where

$$\alpha_1 = \Phi \left( \hat{b} + \frac{\hat{b} + \Phi^{-1}(\alpha/2)}{1 - \hat{a}(\hat{b} + \Phi^{-1}(\alpha/2))} \right), \quad \alpha_2 = \Phi \left( \hat{b} + \frac{\hat{b} + \Phi^{-1}(1 - \alpha/2)}{1 - \hat{a}(\hat{b} + \Phi^{-1}(1 - \alpha/2))} \right),$$

$\hat{b} = \Phi^{-1} \left( \frac{\#(\hat{\beta}_m < \hat{\beta}_0)}{B} \right)$  is the bias-correction,

$$\hat{a} = \frac{\sum_{h=1}^H n_h^{-3} \sum_{i=1}^{n_h} (\hat{\beta}_{hi}^* - \bar{\beta}_h^*)^3}{6 \left( \sum_{h=1}^H n_h^{-2} \sum_{i=1}^{n_h} (\hat{\beta}_{hi}^* - \bar{\beta}_h^*)^2 \right)^{3/2}}$$

is the acceleration (Davis and Hinkley 1997);  $\hat{\beta}_{hi}^*$  is the estimate from the  $i$ th jackknife sample in the  $h$ th strata;  $\bar{\beta}_h^*$  is the mean of the  $n_h$  jackknife estimates from the  $h$ th strata,  $\Phi(\cdot)$  is the CDF and  $\Phi^{-1}(\cdot)$  is the inverse-CDF of the standard normal distribution, and  $\alpha$  is the type I error.

100(1 -  $\alpha$ )% percentile confidence intervals are also calculated where  $(\hat{\beta}_{(\alpha/2)}, \hat{\beta}_{(1-\alpha/2)})$  are the  $\alpha/2$  and  $1 - \alpha/2$  percentiles of the  $B$  estimates. When COMBINE = PRETRANSFORM or COMBINE = FIRST, acceleration is computed as if in a single stratum using  $\hat{\beta}_w$ . An exact two-sided  $p$ -value criterion for covariate imbalance (COVXACT) is calculated as  $\#(|Q_m| \geq |Q_0|)/B$ . While exact analyses are limited to a single outcome in %NParCov4, Program 2.18 in Section 2.9 illustrates how the macro can be repeatedly called to generate exact results for more general problems.

## B SAS Macro %NParCov4

### B.1 General Use and Options

%NParCov4 can be accessed by your program by including the lines:

```
filename nparcov4 "&path";
%include nparcov4(NParCov4) / nosource2;
```

or

```
%include "&path./nparcov4.sas".
```

Data must be set up so there is one line per experimental unit, and there cannot be any missing data. Records with missing data should be deleted or made complete with imputed values prior to analysis. The program does not automatically write output to the SAS listing. The results of individual data sets can be displayed using PROC PRINT. PROC DATASETS will list available data sets generated by %NParCov4 at the end of the SAS log. %NParCov4 has the following options:

1. **OUTCOMES:** List of outcome variables for the analysis. All outcome variables are numeric and at least one outcome is required. Binary or time-to-event outcomes that use **TRANSFORM = LOGISTIC, PODDS, LOGRANK, or WILCOXON** can only have (0,1) values. Here, 1 should indicate that an event occurs, while 0 indicates that an event does not occur (or, in the case of a survival endpoint, that an individual is censored).
2. **COVARS:** List of covariates for the analysis. All covariates must be numeric; categorical variables must be re-expressed using one or more indicator variables. **COVARS** can be left unspecified (blank) for stratified, unadjusted results.
3. **EXPOSURES:** Required numeric exposures for **TRANSFORM = LOGRANK, WILCOXON, or INCDENS**. The same number of exposure variables as **OUTCOMES** is required, and the order of the variables in **EXPOSURES** should correspond to the variables in **OUTCOMES**.
4. **TRTGRPS:** This is a required, single two-level numeric variable that defines the two randomized treatments or groups of interest. Differences are based on higher number (TRT2) - lower number (TRT1). For example, if the variable is coded (1,2), differences computed are treatment 2 - treatment 1.
5. **STRATA:** A single numeric variable that defines strata, this variable can represent strata for cross-classified factors. Specify **NONE** (default) for one stratum (i.e., no stratification). Note that each stratum must have at least one observation per treatment when **HYPOTH = NULL** and at least two observations per treatment when **HYPOTH = ALT**.
6. **C:** Used in calculation of strata weights for all methods.  $0 \leq C \leq 1$  (default = 1). A 0 assigns equal weights to strata, while 1 is for Mantel-Haenszel weights.
7. **COMBINE:** Defines how the analysis will accommodate **STRATA**. Choose one.
  - (a) **NONE** (default) performs analysis as if data are from a single stratum.
  - (b) **LAST** performs covariance adjustment within each stratum, and then takes weighted averages of parameter estimates with weights based on stratum sample size (appropriate for larger strata).
  - (c) **FIRST** obtains a weighted average of treatment group differences (post-transformation, if applicable) across strata, then performs covariance adjustment (appropriate for smaller strata).

- (d) PRETRANSFORM (available for TRANSFORM = LOGISTIC, PODDS, LOGRATIO, or INCdens) obtains a weighted average of treatment group means (pre-transformation) across strata, performs a transformation, and then applies covariance adjustment (appropriate for very small strata).
8. HYPOTH: Determines how variance matrices are computed.
- (a) NULL applies the assumption that the means and covariance matrices of the treatment groups are equal and therefore computes a single covariance matrix for each stratum.
  - (b) ALT computes a separate covariance matrix for each treatment group within each stratum.
9. TRANSFORM: Applies transformations to outcome variables. Choose one.
- (a) NONE (default) applies no transformation and analyzes means for one or more outcome variables as provided.
  - (b) LOGISTIC applies the logistic transformation to the means of one or more (0,1) outcomes.
  - (c) PODDS applies the logistic transformation to multiple means of (0,1) outcomes that together form a single ordinal outcome.
  - (d) LOGRATIO applies the log-transformation based on the natural logarithm to one or more ratios of outcome means.
  - (e) INCdens uses number of events experienced in OUTCOMES and EXPOSURES to calculate and analyze the log-transformation of one or more ratios of incidence density.
  - (f) LOGRANK uses event flags in OUTCOMES (1 = event, 0 = censored) and EXPOSURES to calculate and analyze log-rank scores for one or more time-to-event outcomes.
  - (g) WILCOXON uses event flags in OUTCOMES (1 = event, 0 = censored) and EXPOSURES to calculate and analyze Wilcoxon scores for one or more time-to-event outcomes.
10. ALPHA: Confidence limit for intervals (default = 0.05).
11. EXACT: Provides essentially exact analyses for a single outcome. Choose one.
- (a) NO (default) does not perform resampling analyses;
  - (b) YES performs additional resampling-based analysis. If HYPOTH = ALT, bootstrap resampling is performed with observations sampled with replacement within treatment. If HYPOTH = NULL, permutation resampling is performed with observations sampled without replacement across treatments. This is equivalent to shuffling the treatment codes of the observations.
12. NREPS: Number of random data sets to generate when EXACT = YES (default = 1000).
13. SEED: Random seed when EXACT = YES (default = 0).
14. DSNIN: Input data set. Required.
15. DSNOUT: Prefix for output data sets. Required. For example, if DSNOUT is set to OUTDAT, the following data sets will be created:
- (a) \_OUTDAT\_COVTEST contains the criterion for covariate imbalance. When EXACT = YES and HYPOTH = NULL, the exact criterion for covariate imbalance (COVXACT) is calculated.
  - (b) \_OUTDAT\_DEPTEST contains tests, treatment estimates, and ratios (for TRANSFORM = LOGISTIC, PODDS, LOGRATIO, or INCdens) for outcomes.
  - (c) \_OUTDAT\_CI contains confidence intervals for outcomes when HYPOTH = ALT.
  - (d) \_OUTDAT\_RATIOCI contains confidence intervals for odds ratios (TRANSFORM =

PODDS or LOGISTIC), ratios (TRANSFORM = LOGRATIO), or incidence densities (TRANSFORM = INCDENS) when HYPOTH = ALT.

- (e) \_OUTDAT\_EXACT provided when EXACT = YES
    - i. When HYPOTH = NULL, the data set contains essentially exact, one- and two-sided *p*-values for the outcome.
    - ii. When HYPOTH = ALT, the data set contains percentile, bias-corrected, and accelerated confidence intervals. For TRANSFORM = LOGISTIC, PODDS, LOGRATIO, or INCDENS, the exact percentile, bias-corrected, and accelerated confidence intervals for the ratios are also provided.
  - (f) \_OUTDAT\_HOMOGEN contains the test for homogeneity of odds ratios for TRANSFORM = PODDS.
  - (g) \_OUTDAT\_BETASAMP contains the resampled values of the treatment parameter (BETASAMP). The ratios for TRANSFORM = LOGISTIC, PODDS, LOGRATIO, or INCDENS, are also provided (EXPBETASAMP). The observed treatment parameter is provided (labeled OBSERVED) and for either of the following:
    - i. HYPOTH = NULL, parameters from the permutation data sets are labeled PERMUTATION.
    - ii. HYPOTH = ALT, parameters from bootstrap data sets are labeled BOOTSTRAP, while parameters from jackknife data sets are labeled JACKKNIFE.
  - (h) \_OUTDAT\_SURV contains survival scores for TRANSFORM = LOGRANK or WILCOXON. The scores from this data set from multiple calls of %NParCov4 can be used with TRANSFORM = NONE to analyze log-rank and Wilcoxon scores simultaneously or to analyze survival outcomes with non-survival outcomes.
  - (i) \_OUTDAT\_COVBETA contains the covariance matrix for  $\hat{\beta}$ . The estimates from data set \_OUTDAT\_DEPTEST and the covariance matrix can be used to test hypotheses of linear combinations of the  $\hat{\beta}$ .
  - (j) \_OUTDAT\_STRATABETA contains the strata-specific estimates for  $\hat{\beta}$  when COMBINE = LAST. These estimates can be used to compute acceleration for BC<sub>a</sub> intervals (Appendix A.7) when repeatedly calling %NParCov4 for more general problems.
16. SYMSIZE: Define space requirements for SAS/IML. Should be increased for larger data sets or large NREPS (default = 20000).
17. DETAILS: Print basic analysis information and the PROC DATASETS of analysis results to the SAS log (default = YES).

## B.2 Content of Output Data Sets

Output data sets contain the following variables:

1. \_OUTDAT\_COVTEST
  - (a) TYPE = COVTEST.
  - (b) TRT1 is the value for control (lower value numerically).
  - (c) TRT2 is the value for treatment (higher value numerically).
  - (d) STRATA contains the strata variable.
  - (e) COVARIATES contains the list of covariates.
  - (f) OUTCOMES contains the list of outcomes.
  - (g) EXPOSURES contains the list of exposures for time-to-event endpoints.
  - (h) HYPOTHESIS is either NULL or ALT.

- (i) H is the number of strata levels.
  - (j) DF is the degrees of freedom.
  - (k) Q is the test statistic.
  - (l) PVALUE is the *p*-value.
  - (m) COVXACT is the exact *p*-value when EXACT = YES and HYPOTH = NULL.
2. \_OUTDAT\_DEPTEST.
- (a) TYPE = DEPTEST.
  - (b) TRANSFORM is the transformation applied to the outcomes.
  - (c) TRT1 is the value for control (lower value numerically).
  - (d) TRT2 is the value for treatment (higher value numerically).
  - (e) STRATA contains the strata variable.
  - (f) COVARIATES contains the list of covariates.
  - (g) OUTCOMES contains the list of outcomes.
  - (h) EXPOSURES contains the list of exposures for time-to-event endpoints.
  - (i) HYPOTHESIS is either NULL or ALT.
  - (j) H is the number of strata levels.
  - (k) DF is the degrees of freedom.
  - (l) BETA is the treatment estimate.
  - (m) SEBETA is the standard error of the treatment estimate.
  - (n) RATIO is the exponentiated treatment estimate for TRANSFORM = LOGISTIC, PODDS, LOGRATIO, INCDENS.
  - (o) Q\_J is the test statistic.
  - (p) PVALUE is the *p*-value.
3. \_OUTDAT\_CI.
- (a) TYPE = CI.
  - (b) TRT1 is the value for control (lower value numerically).
  - (c) TRT2 is the value for treatment (higher value numerically).
  - (d) STRATA contains the strata variable.
  - (e) COVARIATES contains the list of covariates.
  - (f) OUTCOMES contains the list of outcomes.
  - (g) EXPOSURES contains the list of exposures for time-to-event endpoints.
  - (h) HYPOTHESIS is either NULL or ALT.
  - (i) H is the number of strata levels.
  - (j) BETA is the treatment estimate.
  - (k) SEBETA is the standard error of the treatment estimate
  - (l) LOWER is the lower limit of the  $100(1 - \alpha)\%$  confidence interval for the treatment estimate.
  - (m) UPPER is the upper limit of the  $100(1 - \alpha)\%$  confidence interval for the treatment estimate.
  - (n) ALPHA is the value of  $\alpha$ .
4. \_OUTDAT\_RATIOCI
- (a) TYPE = RATIOCI.
  - (b) TRT1 is the value for control (lower value numerically).
  - (c) TRT2 is the value for treatment (higher value numerically).
  - (d) STRATA contains the strata variable.

- (e) COVARIATES contains the list of covariates.
  - (f) OUTCOMES contains the list of outcomes.
  - (g) EXPOSURES contains the list of exposures for time-to-event endpoints.
  - (h) HYPOTHESIS is either NULL or ALT.
  - (i) H is the number of strata levels.
  - (j) RATIO is the exponentiated treatment estimate.
  - (k) RATIO\_LOWER is the lower limit of the  $100(1 - \alpha)\%$  confidence interval for the exponentiated treatment estimate.
  - (l) RATIO\_UPPER is the upper limit of the  $100(1 - \alpha)\%$  confidence interval for the exponentiated treatment estimate.
  - (m) ALPHA is the value of  $\alpha$ .
5. \_OUTDAT\_EXACT.
- (a) TYPE = EXACT.
  - (b) TRT1 is the value for control (lower value numerically).
  - (c) TRT2 is the value for treatment (higher value numerically).
  - (d) STRATA contains the strata variable.
  - (e) COVARIATES contains the list of covariates.
  - (f) OUTCOMES contains the list of outcomes.
  - (g) HYPOTHESIS is either NULL or ALT.
  - (h) H is the number of strata levels.
  - (i) TWOSIDED is the two-sided, exact  $p$ -value to test the treatment effect.
  - (j) ONE\_LOWER is the lower one-sided, exact  $p$ -value to test the treatment effect.
  - (k) ONE\_UPPER is the upper one-sided exact  $p$ -value to test the treatment effect.
  - (l) ALPHA is the value of  $\alpha$ .
  - (m) NREPS is the number of resamples.
  - (n) SEED is the value of seed.
  - (o) BIAS is  $\hat{b}$ , the bias used in the calculation of the  $BC_a$  interval.
  - (p) ACCEL is  $\hat{a}$ , the acceleration used in the calculation of the  $BC_a$  interval.
  - (q) ALPHA\_LOW is  $\alpha_1$  used in the calculation of the  $BC_a$  interval.
  - (r) ALPHA\_HI is  $\alpha_2$  used in the calculation of the  $BC_a$  interval.
  - (s) BCA\_LOWER is the lower limit of the  $BC_a$  interval for the treatment effect.
  - (t) BCA\_UPPER is the upper limit of the  $BC_a$  interval for the treatment effect.
  - (u) PCT\_LOWER is the lower limit of the percentile interval for the treatment effect.
  - (v) PCT\_UPPER is the upper limit of the percentile interval for the treatment effect.
  - (w) BCA\_RATIO\_LOWER is the lower limit of the  $BC_a$  interval for the exponentiated treatment effect for TRANSFORM = LOGISTIC, PODDS, LOGRATIO, INCDENS.
  - (x) BCA\_RATIO\_UPPER is the upper limit of the  $BC_a$  interval for the exponentiated treatment effect for TRANSFORM = LOGISTIC, PODDS, LOGRATIO, INCDENS.
  - (y) PCT\_RATIO\_LOWER is the lower limit of the percentile interval for the exponentiated treatment effect for TRANSFORM = LOGISTIC, PODDS, LOGRATIO, INCDENS.
  - (z) PCT\_RATIO\_UPPER is the upper limit of the percentile interval for the exponentiated treatment effect for TRANSFORM = LOGISTIC, PODDS, LOGRATIO, INCDENS.

6. .OUTDAT\_HOMOGEN.
  - (a) **TYPE** = HOMOGEN.
  - (b) **TRT1** is the value for control (lower value numerically).
  - (c) **TRT2** is the value for treatment (higher value numerically).
  - (d) **STRATA** contains the strata variable.
  - (e) **COVARIATES** contains the list of covariates.
  - (f) **OUTCOMES** contains the list of outcomes.
  - (g) **DF** is the degrees of freedom.
  - (h) **BETA** is the treatment estimate.
  - (i) **Q\_C** is the test statistic.
  - (j) **PVALUE** is the *p*-value.
7. .OUTDAT\_BETASAMP.
  - (a) **FLAG** = OBSERVED, PERMUTATION, BOOTSTRAP, JACKKNIFE for type of resampled treatment estimate.
  - (b) **BETASAMP** is the treatment estimate.
  - (c) **EXPBETASAMP** is the exponentiated treatment estimate for **TRANSFORM** = LOGISTIC, PODDS, LOGRATIO, INCDENS.
8. .OUTDAT\_SURV.
  - (a) Same as the input data set but adds survival scores according to **TRANSFORM**. For example, a variable called **VAR** will have one or the other of **LOGRANK\_VAR** or **WILCOXON\_VAR** added to the data set.
9. .OUTDAT\_COVBETA.
  - (a) **TYPE** = COVBETA.
  - (b) Other variables correspond to outcomes and covariates to identify rows and columns of the covariance matrix.
10. .OUTDAT\_STRATABETA.
  - (a) **STRATA\_H** contains the value for a strata level.
  - (b) **BETA\_H** contains the strata-specific treatment estimate.
  - (c) **N\_H** contains the sample size for a strata level.



# Chapter 3

## Dose-Escalation Methods

**Guochen Song (Biogen)**

**Zoe Zhang (Genentech)**

**Nolan Wages (University of Virginia)**

**Anastasia Ivanova (University of North Carolina at Chapel Hill)**

**Olga Marchenko (QuintilesIMS)**

**Alex Dmitrienko (Mediana)**

3.1	Introduction	101
3.2	Rule-based methods	103
3.3	Continual reassessment method	107
3.4	Partial order continual reassessment method	116
3.5	Summary	123
3.6	References	123

This chapter gives an overview of dose-finding methods used in early-phase dose-escalation trials with emphasis on oncology trials. It provides a review of basic dose-escalation designs that use rule-based methods. Special attention is given to model-based methods such as the continual reassessment method for trials with a single agent and its extension (partial-order, continual-reassessment method) for trials with drug combinations. Practical issues related to the implementation of model-based methods are discussed and illustrated using examples from Phase I oncology trials.

### 3.1 Introduction

---

Clinical trials aimed at dose-ranging and dose-finding are conducted at early stages of drug development programs to evaluate the safety and sometimes efficacy profiles of experimental treatments. The general topic of design and analysis of dose-ranging studies that are conducted to test several doses (or dose regimens) of treatments has attracted much attention in the clinical trial literature. This area of clinical trial research is often referred to as dose-response analysis and deals with modeling the relationship between dose and toxicity or clinical response.

The main topic of this chapter is statistical methods used in dose-escalation trials with emphasis on dose-assignment methods arising in oncology settings. Chapter 4 will provide a summary of dose-finding methods commonly used in parallel-group Phase II trials. This includes trend tests, dose-response modeling, and dose-finding strategies.

### 3.1.1 Dose-escalation trials

The first set of clinical trials that are conducted in humans are dose-escalation trials. A very small dose of an experimental treatment is administered to a small number of subjects, e.g., a dose cohort of up to 5 subjects. After this, doses are escalated in a data-driven manner that depends on the safety and pharmacokinetic responses of the patients to provide a thorough characterization of the tolerability profile of the treatment. Such designs are often replicated using repeated administration of a fixed dose of the treatment with the goal of increasing exposure and exploring the range of tolerable doses.

This chapter focuses on statistical methods used in early-phase oncology trials. Most therapeutic-area Phase I clinical trials recruit volunteers rather than patients with a condition of interest. Phase I trials in oncology, however, are conducted in cancer patients, many of whom have run out of available treatment options. The main objective of dose-escalation trials in oncology is to learn about the dose-toxicity relationship of a new therapy. Other objectives are to study pharmacokinetics, i.e., the effect of the body on the treatment, and pharmacodynamics, i.e., the effect of the treatment on the body.

Phase I trials in oncology applications are typically open-label trials with sequential dose-escalation schemes. The primary outcome of most Phase I oncology trials is dose-limiting toxicity (DLT). DLT is typically defined using the National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE). This document is available online at the Cancer Therapy Evaluation Program's website: <http://ctep.cancer.gov/protocolDevelopment/>

The classification scheme introduced in the CTCAE designates laboratory ranges, symptoms, and signs as toxicities or adverse events (AEs) and assigns a grade to each toxicity or AE as follows:

- Grade 1 for a mild AE.
- Grade 2 for a moderate AE.
- Grade 3 for a severe AE.
- Grade 4 for a life-threatening AE.
- Grade 5 for a fatal AE.

DLT is usually defined as a non-hematological toxicity related to the investigational agent of Grade 3 or higher or a hematological toxicity related to the investigational agent of Grade 4 or higher. The probability of DLT is assumed to be a non-decreasing function of the dose. The maximum tolerated dose (MTD) is defined as the dose with a certain probability of the DLT, usually in the 20-25% range. The MTD is used to determine the recommended dose for Phase II trials. Depending on the definition of the DLT, the target range can be as low as 10% and as high as 50%. Because the toxicity profile of the new therapy is yet to be learned, escalation starts with the lowest dose and an adaptive dose-assignment strategy is used with the goal of minimizing the number of patients assigned to low (and hence ineffective) doses and to doses with high likelihood of DLTs. A review of methods used in oncology dose-escalation trials can be found in Rosenberger and Haines (2002) and, more recently, in Sverdlov et al. (2014).

Dose-finding methods in an oncology setting are often classified as *model-based methods* or *non-model-based methods* (also known as *rule-based methods*). Rule-based methods rely on a simple set of assumptions. They only assume that the dose-toxicity curve is non-decreasing with dose. Frequentist, rule-based designs include group up-and-down designs (Wetherill, 1963), cumulative cohort design (Ivanova et al., 2007) and the *t*-statistic design (Ivanova and Kim, 2009). Designs such as the modified toxicity probability interval (mTPI) method from Ji et al.

(2010) and the rapid enrollment method (Ivanova et al., 2016) are examples of Bayesian, rule-based designs. Rule-based methods, including the 3+3 design, group up-and-down designs, and mTPI method, will be discussed in Section 3.2.

Within a model-based framework, the dose-toxicity curve is assumed to follow a parsimonious model. (The number of estimated parameters should be less than the number of doses selected for the trial.) For example, the continual reassessment method (CRM) introduced in O’Quigley et al. (1990) is based on a one- or two-parameter model, and the escalation with overdose control design (Babb, Rogatko, and Zacks, 1998) uses a two-parameter model. The CRM and its extensions have attracted much attention in the literature, and a detailed discussion of the CRM will be presented in Section 3.3. Further, Section 3.4 will introduce the partial order continual reassessment method (POCRM).

The SAS code and data sets included in this chapter are available on the book’s website at <http://support.sas.com/publishing/authors/dmitrienko.html>.

## 3.2 Rule-based methods

---

The key feature of rule-based methods is that patients are assigned to dose levels based on simple pre-defined rules. No assumptions on the dose-toxicity curve are made other than that monotonicity and dose selection decisions are driven by the number of DLTs observed in the current patient cohort. Since the decisions refer to dose escalation or de-escalation, rule-based methods are commonly referred to as *up-and-down methods*. A class of simple up-and-down algorithms, including the popular 3+3 algorithm, is introduced in Section 3.2.1. A Bayesian version of up-and-down methods, known as the modified toxicity probability interval method, is described in Section 3.2.2.

### 3.2.1 Simple up-and-down methods

The most frequently used up-and-down algorithm, Phase I oncology trials is the 3+3 algorithm (Storer, 1989). With this dose-assignment algorithm, patients are assigned in cohorts of three, starting with the lowest dose of study drug with the following provisions:

- If no DLTs occur in the cohort, increase the dose to the next protocol-designated level.
- If 1/3 DLT occurs (meaning that among the three patients, one had a DLT), repeat the same dose in an expanded cohort of six, adding three more patients to receive the study drug at the same original dose.
  - If 1/6 DLT occurs, increase the dose for the next cohort.
  - If  $\geq 2/6$  DLTs occur, stop the trial.
- If  $\geq 2/3$  DLTs occur (meaning that among the three patients, two or more patients had a DLT), stop the trial and assign three more patients to receive the *lower* dose if there are only three patients thus far in that dose level cohort.

The estimated MTD is the dose in which 0/3 or 1/6 patients experienced a DLT. The MTD can be alternatively described as the highest dose level with an observed DLT rate of less than 0.33.

The 3+3 method is easy to understand and implement. The decision rule in this method depends on the number of DLTs observed at the current dose out

of 3 or 6 patients assigned. This approach is attractive to practitioners because of transparent decision rules and an embedded stopping rule. More complex dose-assignment methods, however, can be more efficient in estimating the MTD. In addition, the 3+3 algorithm cannot be used in trials where follow-up for toxicity is long as it might result in a prohibitively long trial.

On average, the 3+3 algorithm selects a dose with the probability of DLT of about 20%. If the target DLT probability is different from 20%, we can modify the cohort size and the decision rule in the 3+3 algorithm to achieve a design that will target the desired quantile. For example, if the goal of a Phase I trial is to select a dose with the probability of toxicity of 10%, the trial's sponsor can use an algorithm where patients are assigned in cohorts of size 5 with decision rules similar to the ones in the 3+3 method. The resulting 5+5 algorithm is defined as follows:

- If no DLTs occur, increase the dose to the next protocol-designated level.
- If 1/5 DLT occurs, repeat the same dose for the next cohort of 5 patients.
  - If 1/10 DLT occurs, increase the dose for the next cohort.
  - If  $\geq 2/10$  DLTs occur, stop the trial.
- If  $\geq 2/5$  DLTs occur, stop the trial and assign five more patients to receive the lower dose if there are only five patients thus far in that dose-level cohort.

Both the 3+3 and the 5+5 methods described above are special cases of the A+B methods (Ivanova, 2006).

A group up-and-down method (Wetherill, 1963) with assignment in cohorts of size three that targets the probability of DLT of about 20% (21% to be precise (Ivanova, 2006)), is defined as follows: Let  $Y$  be the number of toxicities in the most recent cohort of size three. Then the rules are as follows:

- De-escalate the dose if  $Y = 0$ .
- Escalate the dose if  $Y = 1, 2$  or  $3$ .

The process is continued until a prespecified number of patients are assigned (for example, 18 or 24). A change in the decision rule in the group up-and-down method above leads to the change of the target quantile from 21% to 35%. A group up-and-down algorithm that targets the probability of DLT of 35% is defined as follows:

- De-escalate the dose if  $Y = 0$ .
- Repeat the dose if  $Y = 1$ .
- Escalate the dose if  $Y = 2$  or  $3$ .

Ivanova (2006) described how to construct a group up-and-down algorithm that targets a given DLT probability. The cumulative cohort method (Ivanova et al., 2007) and the  $t$ -statistic method (Ivanova and Kim, 2009) are generalizations of a group up-and-down method. The decision rule is based on all patients assigned to the dose, not only on the last cohort of patients.

### **3.2.2 Modified toxicity probability interval method**

The modified toxicity probability interval (mTPI) method was introduced in Ji et al. (2010) as a Bayesian up-and-down algorithm for dose-escalation oncology trials. Unlike the frequentist up-and-down method, this method relies on simple yet flexible

decision rules and, for this reason, has attracted much attention in Phase I trials. For more information on applications of the mTPI algorithm and its extensions, see Ji and Wang (2013) and Guo et al. (2017). Theoretical properties of the mTPI method were studied in Ji et al. (2010), and it was shown that this method exhibits optimal performance based on the posterior expected loss.

The mTPI method can be thought of as a Bayesian extension of A+B designs. The mTPI rules are based on an intuitive approach aimed at assessing the likelihood of overdosing and underdosing given the data in the current patient cohort. To define the rules, consider a dose-escalation trial with a pre-defined set of doses defined by  $x_1, \dots, x_k$ . Suppose that  $n_i$  patients are assigned to the  $i$ th cohort and  $m_i$  patients experience a DLT (here  $0 \leq m_i \leq n_i$ ). Let  $p^*$  denote the target probability of toxicity in the trial. Consider the following partitioning scheme with three toxicity intervals:

- Underdosing interval,  $[0, p^* - \varepsilon_1]$ .
- Equivalence interval,  $[p^* - \varepsilon_1, p^* + \varepsilon_2]$ .
- Overdosing interval,  $(p^* + \varepsilon_2, 1]$ .

Here,  $\varepsilon_1$  and  $\varepsilon_2$  are pre-specified positive constants that define the equivalence interval and help quantify the uncertainty around the selected value of  $p^*$ . It is common to set  $\varepsilon_1$  and  $\varepsilon_2$  to 0.05.

Using the observed number of DLTs in the current cohort, the trial's sponsor can easily compute the posterior probabilities of the toxicity intervals as well as the *unit probability mass* (UPM) of each interval. The unit probability mass of an interval is equal to its posterior probability divided by the interval's length and thus the unit probability masses of the underdosing, equivalence, and overdosing intervals are given by

$$\begin{aligned} \text{UPM}_U &= \frac{P(p_i < p^* - \varepsilon_1 | D)}{p^* - \varepsilon_1}, \\ \text{UPM}_E &= \frac{P(p^* - \varepsilon_1 \leq p_i \leq p^* + \varepsilon_2 | D)}{\varepsilon_1 + \varepsilon_2}, \\ \text{UPM}_O &= \frac{P(p_i > p^* + \varepsilon_2 | D)}{1 - p^* - \varepsilon_2}, \end{aligned}$$

respectively, where  $D$  denotes the observed data and  $p_i$  denotes the true toxicity probability in the current cohort. Assuming a uniform prior for  $p_i$ , it is easy to show that the posterior distribution of  $p_i$  is a beta distribution with the shape parameters given by  $1 + m_i$  and  $1 + n_i - m_i$ . Thus the unit probability masses are equal to

$$\begin{aligned} \text{UPM}_U &= \frac{F(p^* - \varepsilon_1 | 1 + m_i, 1 + n_i - m_i)}{p^* - \varepsilon_1}, \\ \text{UPM}_E &= \frac{F(p^* + \varepsilon_2 | 1 + m_i, 1 + n_i - m_i) - F(p^* - \varepsilon_1 | 1 + m_i, 1 + n_i - m_i)}{\varepsilon_1 + \varepsilon_2}, \\ \text{UPM}_O &= \frac{1 - F(p^* + \varepsilon_2 | 1 + m_i, 1 + n_i - m_i)}{1 - p^* - \varepsilon_2}, \end{aligned}$$

where  $F(x|\alpha, \beta)$  is the cumulative distribution function of the beta distribution with the shape parameters  $\alpha$  and  $\beta$ .

The resulting dose-assignment rule chooses the dose for the next patient cohort that corresponds to the highest unit probability mass. For example, if the highest UPM corresponds to the overdosing interval, the most sensible decision will be to reduce the dose. Similarly, a decision to increase the dose is made if the highest UPM corresponds to the underdosing interval. Finally, patients in the next cohort will be

dosed at the same level if the equivalence interval is the most likely interval based on the UPM. This decision is conceptually similar to the decision to expand the current cohort if exactly one out of three patients has a DLT in the 3+3 algorithm.

In addition to these dose-assignment rules, the trial's sponsor may want to consider a rule for excluding doses based on an excessive probability of toxicity. Let  $\lambda$  denote a pre-defined threshold for the posterior probability of unacceptably high toxicity, e.g.,  $\lambda = 0.95$ . A decision to de-escalate will be made, and the current dose will be excluded from the set of eligible doses if

$$P(p_i > p^* | D) \geq \lambda.$$

As a quick comment, the mTPI method essentially relies on a basic Bayesian hierarchical model with independent prior distributions for the toxicity probabilities across the dose levels, i.e.,  $p_1, \dots, p_k$ . For this reason, posterior probabilities of the toxicity intervals are computed using the data from the current cohort only. They disregard the toxicity information observed earlier in the trial. The standard mTPI algorithm can be extended to the case of dependent priors. However, Ji et al. (2010) argued that the use of independent prior is more appropriate in small Phase I trials.

Finally, to determine the MTD, Ji et al. (2010) proposed to apply an isotonic transformation to the posterior means of the toxicity probabilities in each cohort. The resulting adjusted estimates of the cohort-specific probabilities are monotonically non-decreasing. The MTD is defined as the dose for which the adjusted toxicity probability is closest to the target probability  $p^*$ .

To illustrate the mTPI algorithm, Program 3.1 performs the Bayesian calculations to derive the dose-assignment rules in a cohort of five patients. The target probability of toxicity in the trial is set to  $p^* = 0.3$ , and the equivalence interval is defined based on  $\varepsilon_1 = \varepsilon_2 = 0.05$ . The posterior probability of unacceptably high toxicity is computed using  $\lambda = 0.95$ . Given the number of observed DLTs in the cohort, denoted by  $m$ , the program computes the unit probability masses for the underdosing, equivalence, and overdosing intervals (`upmu`, `upme`, `upmo`), as well as the posterior probability of excessive toxicity (`stopprob`).

### **PROGRAM 3.1    mTPI-based decision rules with a cohort of size five**

```

data mtpi;
length decision $2;
n = 5;

probtox = 0.3;
epsilon1 = 0.05;
epsilon2 = 0.05;

do m = 0 to n;
    prob1 = 1 - cdf('beta', probtox + epsilon2, 1 + m, 1 + n - m);
    prob2 = cdf('beta', probtox - epsilon1, 1 + m, 1 + n - m);
    stopprob = 1 - cdf('beta', probtox, 1 + m, 1 + n - m);
    upmu = prob2 / (probtox - epsilon1);
    upmo = prob1 / (1 - probtox - epsilon2);
    upme = (1 - prob1 - prob2) / (epsilon1 + epsilon2);
    * De-escalate;
    if upmo > upmu & upmo > upme then do;
        if stopprob < 0.95 then decision = "D";
        * De-escalate and remove this dose;
        else decision = "DR";
    end;
    * Escalate;
    if upmu > upmo & upmu > upme then decision = "E";

```

```

* Stay at the same dose;
if upme > upmo & upme > upmu then decision = "S";
output;
end;

keep m upmu upme upmo stopprob decision;

run;

proc print data = mtpi noobs;
var m upmu upme upmo stopprob decision;
run;

```

---

	m	upmu	upme	upmo	stopprob	decision
0	3.28809	1.02560	0.11603	0.11765	E	
1	1.86426	2.14856	0.49089	0.42018	S	
2	0.67773	1.83481	0.99552	0.74431	S	
3	0.15039	0.79826	1.35781	0.92953	D	
4	0.01855	0.17683	1.50412	0.98907	DR	
5	0.00098	0.01594	1.53563	0.99927	DR	

---

The decisions are summarized in Output 3.1. If no one experiences a DLT ( $m = 0$ ), the highest unit probability mass is associated with the underdosing interval, which leads to a decision to escalate (E). With  $m = 1$  or  $m = 2$ , the UPM for the equivalence interval is greater than the other two UPMs. Thus, the most appropriate decision is to stay at the current dose (S). When the number of DLTs is greater than 2, a de-escalation decision is made since the highest unit probability mass is now associated with the overdosing interval (D). Further, the posterior probability of excessive toxicity is greater than the prespecified threshold if  $m = 4$  or  $m = 5$  and the associated decision is to de-escalate the dose (D) and remove this particular dose from consideration to make sure it will never be examined in the trial again (R). As shown in Output 3.1, the mTPI method is very easy to implement in Phase I trials since, just like other rule-based methods for setting up dose-escalation designs, this method relies on a set of predefined decision rules that depend only on the cohort size and number of DLTs observed in a cohort.

### 3.3 Continual reassessment method

---

This section provides a brief overview of the continual reassessment method (CRM) and illustrates its implementation in SAS using a Phase I trial example motivated by a study with a single agent. The CRM is an efficient dose-finding method for early-phase oncology trials because it uses all available information, not only information at the current dose. Another advantage of the CRM is that it can be modified for use in more complex dose-finding problems. For example, the TITE-CRM (Cheung and Chappell, 2000) is a CRM modification for trials with long follow-up for toxicity. The partial order continual reassessment method (POCRM) introduced in Wages, Conaway, and O'Quigley (2011) is an extension of the CRM to dose-escalation trials with drug combinations where the goal is to identify a dose combination or combinations that yield a certain probability of DLT.

The CRM requires specification of the working model and a prior on the model parameters. It also requires a computer program to implement. The decision process is not as transparent as with the rule-based methods introduced in Section 3.2.

These are the barriers to a wider adoption of the CRM in early-phase oncology trials. In this section, we give advice on how to implement the CRM and POCRM in practice and provide the SAS code for implementation of these methods. The reader can find more information on the CRM and its modifications, e.g., BMA-CRM (Yin and Yuan, 2009), TITE-CRM (Chueng and Chappell, 2000), and Zink and Menon, 2016.

**EXAMPLE: Case study 1 (Phase I trial in patients with advanced HER2-expressing solid tumors)**

The following case study will be used in this section to motivate and illustrate dose-finding methods based on the CRM. Consider a Phase I trial in patients with advanced HER2-expressing solid tumors. The goal is to estimate the MTD of the study drug administered every 3 weeks. Six dose levels are selected for the trial. The DLT assessment period is 3 weeks after the first dose administration. The MTD is defined as the dose level with the probability of toxicity of 25%. It is planned to assign 21 patients in cohorts of size 3. If the MTD is within the range of doses selected for the trial, simulations with different assumptions showed that with 21 patients, the correct dose or the dose adjacent to the correct dose would be most likely to be selected as the MTD.

### 3.3.1 Modeling frameworks

Consider a Phase I trial with  $k$  dose levels denoted by  $x_i$ ,  $i = 1, \dots, k$ , and let  $D = \{(y_i, n_i), i = 1, \dots, k\}$  represent the study data, where  $y_i$  is the number of toxicities observed among the  $n_i$  patients assigned to the dose  $x_i$ . The trial is conducted to identify the dose level  $x_i$  with the probability of toxicity  $\pi(x_i)$  closest to the target probability  $\theta$ .

The binary toxicity response of a patient assigned to the dose level  $x_i$  is postulated as a Bernoulli variable with the toxicity probability  $\pi(x_i)$ , where  $\pi(x_i)$  is a nondecreasing function of  $x_i$ . The likelihood of the observed data is a product of binomial likelihoods with the parameters  $(n_i, y_i)$ .

Within a model-based framework, a dose-toxicity model is assumed, i.e.,

$$\pi(x, \beta) \equiv F(x, \beta),$$

where  $\beta$  is the model parameter that is to be sequentially estimated from the accumulated data and  $x$  is the nominal dose level.

## CRM models

Two most frequently used models in the CRM literature are the power model

$$F(x, \beta) = x^\beta \quad \text{for } 0 < x < 1$$

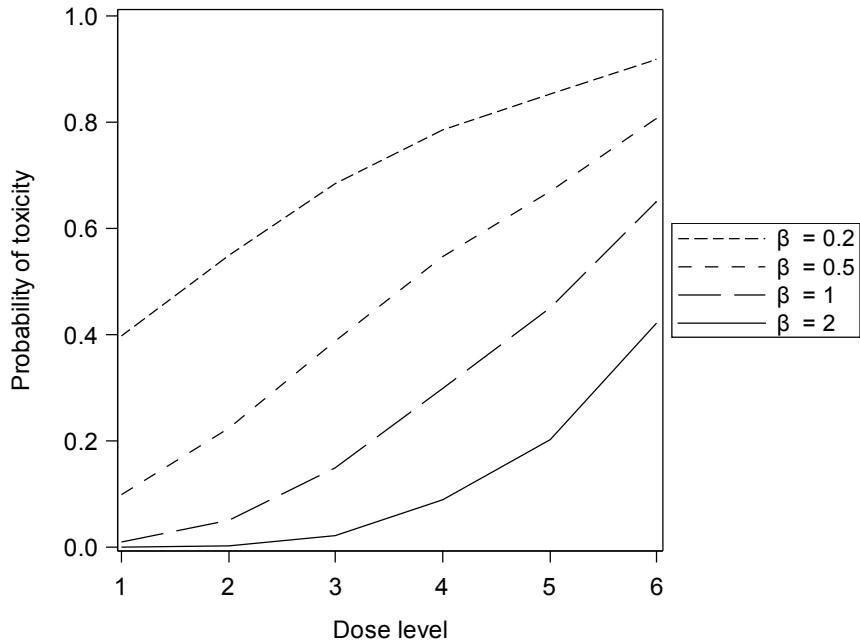
and the one-parameter logistic model

$$F(x, \beta) = \frac{\exp(a_0 + \beta x)}{1 + \exp(a_0 + \beta x)} \quad \text{for } -\infty < x < \infty$$

where the intercept  $a_0$  is fixed and  $\beta$  is to be estimated. Examples of the power model and the one-parameter logistic model with a fixed intercept are shown in Figures 3.1 and 3.2.

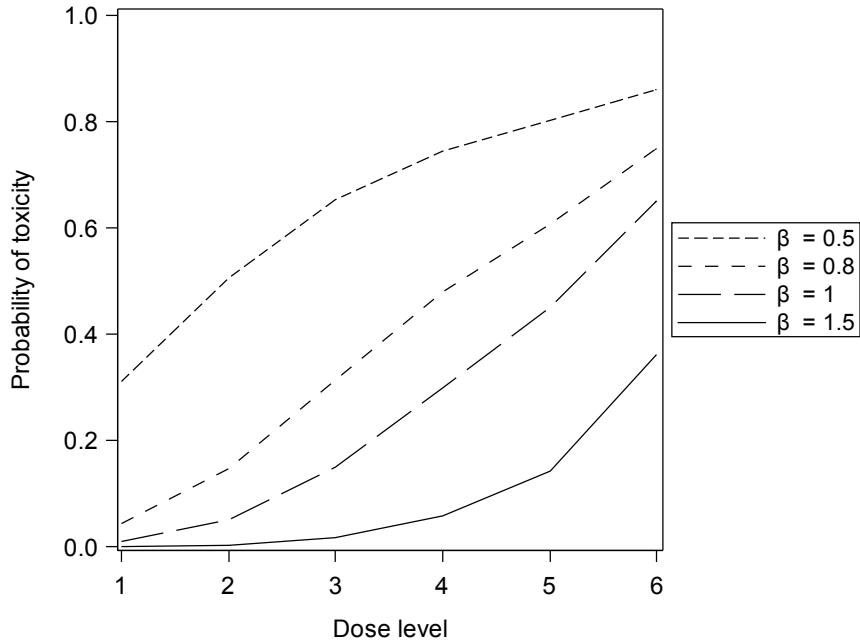
With the parameterization above, the parameter  $\beta$  in these models is restricted to positive values to ensure an increasing dose-toxicity function in the dose level

**Figure 3.1**  
Power model



Power model with  $\beta = 0.2, 0.5, 1$ , and  $2$ ,  $x = 0.01, 0.05, 0.15, 0.30, 0.45$ , and  $0.65$ .

**Figure 3.2**  
Logistic model



One-parameter logistic model with the fixed intercept  $a_0 = 3$  and  $\beta = 0.5, 0.8, 1$ , and  $1.5$ ,  $x = 0.01, 0.05, 0.15, 0.30, 0.45$ , and  $0.65$ .

$x$ . This positivity constraint on  $\beta$  could present some difficulties in estimation in the presence of small samples that are typically encountered in dose-escalation trials. Alternatively, the parameter  $\beta$  is not restricted to  $(0, \infty)$  if the following parameterization is used in the two models:

$$F(x, \beta) = x^{\exp(\beta)}, \quad 0 < x < 1, \quad (3.1)$$

$$F(x, \beta) = \frac{\exp(a_0 + \exp(\beta)x)}{1 + \exp(a_0 + \exp(\beta)x)}, \quad -\infty < x < \infty, \quad (3.2)$$

where  $\beta \in (-\infty, \infty)$ .

It is worth mentioning that the CRM does not require  $F$  to be a correct model for  $\pi$ . It only requires that  $F(x, \beta)$  be strictly increasing in the dose level  $x$ . A one-parameter CRM is under-parameterized and is unlikely to produce a correct fit to the dose-toxicity curve over the entire spectrum of doses. However, as long as the proposed model approximates the relationship reasonably well around the target dose, it enables us to make the correct decision about the MTD.

## Bayesian CRM

The original formulation of the CRM (O’Quigley et al., 1990) uses a Bayesian approach. The model parameter  $\beta$  is assumed to be random and the uncertainty in the chosen dose-toxicity model is expressed by a prior distribution of the model parameter, denoted as  $g(\beta)$ . Using the parameterization defined in Equations (3.1) and (3.2), it is common to employ a normal distribution with mean 0 for  $g(\beta)$ . If  $\beta$  is restricted to be positive, the prior for  $\beta$  is a unit exponential distribution.

If  $\mathcal{D}$  denotes the data after  $j$  patients were enrolled and their toxicity data observed, the likelihood function is given by

$$\mathcal{L}(\mathcal{D} | \beta) \propto \prod_{i=1}^k \{F(x_i, \beta)\}^{y_i} \{1 - F(x_i, \beta)\}^{n_i - y_i}, \quad (3.3)$$

where  $F(x, \beta)$  is the dose-toxicity model. The density function of the posterior distribution of  $\beta$  is updated using the following Bayesian calculation:

$$f(\beta | \mathcal{D}) = \frac{\mathcal{L}(\mathcal{D} | \beta) g(\beta)}{\int \mathcal{L}(\mathcal{D} | \beta) g(\beta) d\beta}. \quad (3.4)$$

The posterior mean of the toxicity probability for  $i$ th dose, given the accumulated information is:

$$\hat{\pi}(x_i) = \int F(x_i, \beta) f(\beta | \mathcal{D}) d\beta. \quad (3.5)$$

An alternative estimate for this updated toxicity probability, sometimes referred to as the “plug-in” estimate, is:

$$\tilde{\pi}(x_i) = F(x_i, \tilde{\beta}), \quad (3.6)$$

where

$$\tilde{\beta} = \int \beta f(\beta | \mathcal{D}) d\beta$$

is the posterior mean of the parameter  $\beta$  based on the data from the first  $j$  patients. The next patient is assigned to the dose with the estimated probability of DLT closest to the target level.

## CRM working model

It is important to note that constants  $x_i$ , often referred to as the *skeleton* of the model, do not correspond to the actual amount of drug, e.g. mg/m<sup>2</sup>. The skeleton does not have to be related to the actual doses or the probabilities of DLTs at the actual doses. Rather, it is selected to yield good operating characteristics of the CRM as described in Lee and Cheung (2009). To accomplish this, the following values need to be specified:

- Half-width of the indifference interval ( $\delta$ ).
- Target toxicity probability ( $\theta$ ).
- Prior belief of MTD location (dose level  $\nu_0$ ).
- Number of doses ( $k$ ).
- Dose-toxicity function ( $F$ ).

### 3.3.2 Implementation of CRM

In most Phase I trials, the CRM is implemented with the modifications proposed in Faries (1994), Goodman (1995), and Piantadosi et al. (1998). The modified CRM algorithm is given in this way:

1. Start with the lowest dose level.
2. Avoid skipping dose levels during dose escalation.
3. Assign patients in cohorts rather than one patient at a time.

To implement the Bayesian CRM in practice, the following steps should be taken:

1. Select a dose-toxicity model  $F(x_i, \beta)$ , where  $x_i, i = 1, \dots, k$ , are the pre-determined dose levels and  $\beta$  is a model parameter.
2. Select a prior distribution for  $\beta$ .
3. Select the target probability of toxicity,  $\theta$  (for example,  $\theta = 0.25$ ).
4. Select a working model (skeleton of the CRM) using the `%GetPrior` macro introduced below.
5. Select a cohort size.
6. Assign the first cohort of patients to the dose level  $x_1$ . Once the dichotomous responses  $Y_j$  are observed, the posterior distribution of  $\beta$  is updated using Equation 3.4. Then the probability of toxicity  $\hat{\pi}(x_i)$  at each dose level  $x_i$  is updated using Equation 3.5.
7. Assign the next cohort of patients to the dose level  $x_i$  with  $\hat{\pi}(x_i)$  closest to the target probability  $\theta$  with the restriction that no skipping of dose level during dose escalation is allowed.
8. Continue the process until a predetermined number of patients is assigned.

Various rules that guide the total number of patients enrolled in the trial have been proposed for the CRM. For example, it might be desirable not to stop the trial unless a certain number of patients have been assigned to the estimated MTD or if there is a high probability that including, say, three more patients will lead to recommending a different dose (O'Quigley and Reiner, 1998). Alternatively, one can stop the trial early if the recommended dose will not change if several subsequent cohorts of new patients are enrolled. One should keep in mind, however, that the

fact that the recommended dose do not change when several cohorts are assigned to that dose does not mean that the true MTD has been found. This is because each result is an observation of the underlying random variable, and the true MTD can only be estimated.

## CRM in Case study 1

The Phase I clinical trial from Case study 1 will be used to illustrate the CRM methods in practice. Let us assume that the power function for the dose-toxicity model is given by

$$F(x_i, \beta) = x_i^{\exp(\beta)},$$

where  $x_i$ ,  $i = 1, \dots, 6$ , are the predefined skeleton values, and  $\beta$  is a model parameter. The prior distribution for  $\beta$  is assumed to be  $N(0, 1.34)$ .

The %GetPrior macro computes the skeleton given trial-specific parameters. The macro has the following parameters:

- **delta** is the half-width of the indifference interval.
- **target** is the target toxicity probability.
- **mdt0** specifies the prior belief of the MTD location.
- **nlevel** is the total number of doses.
- **model** is the selected dose-toxicity model, i.e., power model (**POWER**) or logistic model (**LOGISTIC**).

Based on the results in Lee and Cheung (2009), Cheung (2013) recommends using  $\delta = 0.25 \times \theta$  and sets the expected MTD at the middle dose if no information is available. If the expected MTD is set at the lowest dose, escalation will be slow—that is, many patients without DLTs will be required to be assigned to each dose before the dose can be escalated.

The skeleton of the CRM can be determined in Case study 1 using the %GetPrior macro as shown in Program 3.2.

### PROGRAM 3.2    Skeleton of the CRM

```
%GetPrior(delta=0.1, target=0.25, mdt0=3, nlevel=6, model=POWER)
```

Output 3.2 lists the resulting skeleton.

---

**Output from  
Program 3.2**

---

0.0108127 0.081663 0.25 0.4643377 0.6540842 0.7906349

---

To implement the CRM in the Phase I trial, we will assign the first cohort of three patients to the dose level  $x_1$  and observe the toxicity response. According to the prior distribution and the data observed, the model is updated, and the mean probability of toxicity at each dose level is obtained. The implementation of this step relies on the %CRM macro, which implements the general CRM algorithm. Here is the list of the macro parameters:

- **sim** is the number of simulations runs in the simulation mode (default value is 1,000). This parameter is ignored if **realdata=TRUE**.
- **seed** is the seed for the simulation runs.

- `truetox` is the true toxicity rate for each level used to generate data.
- `startdose` is the starting dose in the simulation mode (default value is 1). This parameter is ignored if `realdata=TRUE`.
- `curdose` is the current dose in the data analysis mode. This parameter is ignored if `realdata=FALSE`.
- `cohortsiz`e is the size of each patient cohort (default value is 3).
- `maxcohort` is the maximum number of cohorts. Note that the product of `maxcohort` and `cohortsiz`e is the total number of patient) (default value is 10).
- `outf` is the output file (default is OUTF).
- `postdirect` determines if the plug-in method used or not (default is TRUE).
- `target` is the targeted toxicity (default value is 0.25).
- `model` is the selected dose-toxicity model, i.e., power model (POWER) or logistic model (LOGISTIC). The default value is POWER.
- `priors2` is the variance for the normal prior (default value is 1.34).
- `alpha0` is the intercept for the logistic model (default value is 3).
- `prior` is the prior distribution, i.e., unit exponential (UNITEXP), normal (NORMAL) or flat (FLAT). The default value is NORMAL.
- `restrict` determines if dose skipping is allowed (default value is FALSE).
- `best` specifies the dose selection strategy. The selected dose is closest to the target in absolute value if `best=BEST` and the selected dose does not exceed the target value if `best=SAFE`.
- `nstop` specifies the sample size at which the trial stops (default value is 31).
- `realdata` determines the macro's mode (data analysis mode or simulation mode). The macro runs once and analyzes the real clinical data in the `trialdata` data set if `realdata=TRUE`. Note that the simulation settings are ignored in this case.

The %CRM macro is invoked in this particular example as shown in Program 3.3.

### PROGRAM 3.3 CRM implementation

```
%CRM(realdata=TRUE, trialdata=dinput, curdose=1,postdirect=FALSE);
```

As an illustration, the output of Program 3.3 is presented in Table 3.1. The table lists the results when no DLT is observed in the first cohort with the mean posterior probability of toxicity (`postdirect=FALSE`). The dose level recommended for the next patient cohort would be the dose level 3 if the level with the toxicity probability closest to the target probability of 25% is to be selected. However, because skipping of dose levels is not allowed, the dose level selected for the next cohort is the dose level 2.

**TABLE 3.1 Output of Program 3.3**

Enrolled	Toxicity observed	Prior toxicity probability	Toxicity probability	Estimated parameter	Next dose
3	0	0.011	0.04	1.440	3
0	0	0.082	0.105	.	.
0	0	0.25	0.215	.	.
0	0	0.464	0.361	.	.
0	0	0.654	0.518	.	.
0	0	0.791	0.662	.	.

We continue the process until the seventh cohort has been treated at the recommended dose level 5 for the total of 21 patients. The trial then will be stopped and

**TABLE 3.2 Mean probability of toxicity and dose level recommended by the CRM algorithm**

Cohort	Dose	Tox	Mean P(Toxicity)						Selected dose
			1	2	3	4	5	6	
Prior			0.011	0.082	0.25	0.464	0.654	0.791	
1	1	0/3	0.002	0.027	0.136	0.331	0.542	0.713	2
2	2	0/3	0.000	0.008	0.068	0.226	0.440	0.635	3
3	3	0/3	0.000	0.001	0.025	0.129	0.322	0.535	4
4	4	0/3	0.000	0.000	0.005	0.054	0.198	0.409	5
5	5	0/3	0.000	0.000	0.000	0.013	0.091	0.266	6
6	6	1/3	0.000	0.000	0.001	0.024	0.127	0.320	6
7	6	3/3	0.000	0.000	0.015	0.097	0.275	0.490	5
8	5	0/3	0.000	0.000	0.007	0.063	0.216	0.429	5
9	5	1/3	0.000	0.000	0.009	0.075	0.239	0.454	5
10	5	0/3	0.000	0.000	0.005	0.055	0.202	0.413	5

Note: Selected dose is the dose level selected for the next cohort.

the estimated MTD is defined as the dose level that would have been assigned to the next patient. Using the %CRM macro in a simulation mode, Program 3.4 generates ten different scenarios for this design but with a total of 30 patients and with the plug-in method (with postdirect=TRUE).

#### PROGRAM 3.4 CRM simulation

```
%CRM(sim=10, model=POWER, postdirect=FALSE, maxcohort=10);
```

Table 3.2 presents a summary of the information produced by Program 3.4 for each cohort and shows one possible scenario among the simulations. The cohorts 8-10 show that, if the trial was planned for 30 patients instead of 21, it would treat the last three cohorts with the MTD dose level 5.

The SAS code presented above allows us to select from the power model and the one parameter logistic model as well as prior distribution for the parameter  $\beta$ . The plugged-in method and the method that relies on the posterior mean probability of toxicity were both implemented through numerical integration using the QUAD subroutine in PROC IML. To implement this function, it is sometimes necessary to provide the peak and scale parameters. We used a nonlinear optimizer (NLPTR subroutine in PROC IML) to find the peak of the posterior marginal distribution.

The unit exponential distribution has been used as a prior distribution with the power model  $F(x, \beta) = x^\beta$  in most published examples and is selected as one of the options to implement in the SAS code. A normal prior with mean 0 and standard deviation  $\sigma_\beta$  is also often used with the model  $F(x, \beta) = x^{\exp(\beta)}$ . A most common choice for  $\sigma_\beta$  is 1.34, but it can be calibrated to obtain the least informative normal prior. See, for example, Cheung (2011, Chapter 9).

We have also implemented the maximum likelihood CRM (O'Quigley and Shen, 1996) with the power model by specifying a flat prior on the parameter  $\beta$ . It is worth mentioning that the MCMC procedure can be used to calculate the posterior probability as well. For example, the power model with a normal prior used in the clinical trial example above can be analyzed using Program 3.5 after three patients enrolled in the first cohort and no toxicity was observed.

#### PROGRAM 3.5 CRM simulation using MCMC

```
data dinput;
input doselevel p0k y n;
cards;
1 0.011 0 3
```

```

2 0.082 0 0
3 0.250 0 0
4 0.464 0 0
5 0.654 0 0
6 0.791 0 0
;
run;

proc mcmc data=dinput nmc=50000 seed=27707 stats=summary outpost=outp
monitor=(pp1 pp2 pp3 pp4 pp5 pp6);
parm a;
array sk(6) 0.011 0.082 0.250 0.464 0.654 0.791;
array pp(6);
prior a ~ normal(0, var=1.34);
beta=exp(a);
do i=1 to 6;
pp[i]=sk[i]**beta;
end;
p=p0k**beta;
model y ~ binomial(n,p);
run;

```

---

**Output from Program 3.5**
**Posterior Summaries**

Parameter	N	Mean	Standard Deviation	Percentiles		
				25%	50%	75%
pp1	50000	0.0382	0.0797	0.000001	0.00213	0.0354
pp2	50000	0.1029	0.1432	0.00133	0.0330	0.1568
pp3	50000	0.2128	0.2076	0.0254	0.1510	0.3581
pp4	50000	0.3585	0.2494	0.1309	0.3510	0.5662
pp5	50000	0.5165	0.2554	0.3248	0.5604	0.7301
pp6	50000	0.6619	0.2286	0.5375	0.7264	0.8406

---

Output 3.5 shows that the posterior mean of the third dose is 0.2128, which is fairly close to the value shown in Output 3.3 (0.215). The code in Program 3.5 can be updated to support a one-parameter logistic model by replacing the power function with the logistic function in calculating posterior probabilities. Models with two parameters, as well as EWOC models, can also be implemented as the procedure creates the posterior samples for any variable of interest using the MONITOR option.

### Implementation of Bayesian two-parameter logistic models

Storer (1989) studied the estimation of the MTD using a two-parameter logistic regression model that is fitted to the toxicity data at the conclusion of a trial. Storer (1989) noted that, although the two-parameter logistic model was conceptually reasonable, it could pose computational difficulties for small samples, producing slope estimates that are less than or equal to 0 or infinity. For a set of assumed dose-toxicity curves, the estimated slope of the logistic curve based on data from the 3+3 algorithm “does not give convergent estimates a usefully large fraction of the time” (Storer, 1989). Although this was not the focus in conducting his study,

his results suggest that factors such as the sample size, true dose-toxicity curve, and design could create computational difficulties for maximum likelihood estimation in the two-parameter logistic model. In considering a Bayesian approach to designing Phase I trials using a two-parameter logistic model, Gatsonis and Greenhouse (1992) noted that, in order to have a proper posterior distribution, until the first DLT is observed, the prior for the MTD needs to be informative.

Some cautions need to be taken when using the two-parameter logistic model. Cheung (2011) demonstrated that the two-parameter logistic model can be “rigid” in that the additional flexibility of the two-parameter model confines the estimation to suboptimal doses. He illustrated this idea of rigidity with an example of no DLTs out of  $n_1$  patients on the dose level  $x_1$  and one DLT out of 3 patients on the dose level  $x_2$ . With a target of  $\theta = 0.10$ , the maximum likelihood estimates for the two-parameter logistic model will always lead to a recommendation of  $x_1$ , regardless of what outcomes are observed at  $x_1$ . It is possible for this rigidity to be overcome by an appropriate choice of priors in the Bayesian framework, which highlights the importance of the prior in the two-parameter logistic setting, as pointed out by Gatsonis and Greenhouse (1992).

The use of the two-parameter model has been recommended in recent papers, despite the problems inherent in the model. Neuenschwander, Branson, and Gsponer (2008) reported on a Phase I trial in which it appeared that the one-parameter CRM called for escalation after two patients in a cohort of size 2 both experienced DLTs. The authors conjectured that this behavior was due to the use of the one-parameter power model. But it has been pointed out that the behavior is more likely due to a poor choice of skeleton (Iasonos and O’Quigley, 2014); an overly informative prior that puts too much prior probability on high dose levels being the MTD (Iasonos and O’Quigley, 2012); or deviations from the original design that allowed investigators to skip two dose levels. For further discussion on the small- and large-sample impact of over-parameterization in the CRM, we refer the reader to Iasonos et al. (2016).

### **3.4 Partial order continual reassessment method**

---

Recently, there has been an increasing interest in investigating the potential of drug combinations for patient treatment. The motivation to treat with drug combinations stems from the desire to improve the response in patients, especially those who have been resistant to traditional treatment. Multi-agent dose-escalation trials present the significant challenge of identifying an MTD combination or combinations of the agents being tested with the typically small sample sizes involved in Phase I trials. In this section, the partial order continual reassessment method (POCRM) with applications to drug-combinations studies is discussed and illustrated using a published drug-combination trial.

**EXAMPLE: Case study 2 (Phase I trial in patients with advanced cancer)**

As previously noted, a key assumption used in single-agent dose-escalation trials is the monotonicity of the dose-toxicity curve. In this case, the curve is said to follow a *simple order* because the ordering of DLT probabilities for any pair of doses is known, and administration of greater doses of the agent can be expected to produce DLTs in increasing proportions of patients. In studies testing combinations of agents, the probabilities of DLT associated with the dose combinations often follow a *partial order* in that there are pairs of dose combinations for which the ordering of the probabilities is not known. As an example of a partial order, consider a Phase I trial of BMS-214662 in combination with paclitaxel and carboplatin in patients with advanced cancer (Dy et al., 2005). The trial investigated escalating doses of

BMS-214662 in combination with escalating doses of paclitaxel, along with a fixed dose of carboplatin. The six dose levels based on the combination of BMS-214662 and paclitaxel tested in this trial are listed in Table 3.3.

**TABLE 3.3 Treatment labels for dose combinations in Case study 2**

Treatment	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
BMS-214662	80	80	120	160	160	225
Paclitaxel	135	175	175	175	225	175

## Partial order

---

Considering the dose levels defined in Table 3.3, as we move from  $x_1$  to  $x_5$ , the dose of one agent remains fixed, while the dose of the other agent is escalated. Therefore, in terms of the probability of DLT at each combination, it is reasonable to assume that

$$\pi(x_1) < \pi(x_2) < \pi(x_3) < \pi(x_4) < \pi(x_5).$$

For  $x_5$  and  $x_6$ , the dose of BMS-214662 is escalated, while the dose of Paclitaxel is de-escalated. It might not be reasonable to assume that  $x_5$  is less toxic than  $x_6$ . It could be that

$$\pi(x_5) < \pi(x_6) \text{ or } \pi(x_6) < \pi(x_5),$$

which creates a partial order.

A traditional approach to this problem is to prespecify an escalation path (i.e., guess an ordering) and apply a single-agent trial design along this path. The disadvantage of this approach is that it could produce highly misleading results if the assumed ordering is incorrect. Furthermore, in larger dimensional problems, it could limit the number of combinations that can be considered, potentially missing promising dose combinations located outside the path.

Rather than work with a single ordering, another approach to deal with this added complexity is to specify multiple possible orderings and appeal to established model selection techniques. Taking into account these known and unknown relationships between the combinations in the clinical trial example, we can formulate two possible orderings, indexed by  $m$ , of the toxicity profile:

$$m = 1 : \pi(x_1) < \pi(x_2) < \pi(x_3) < \pi(x_4) < \pi(x_5) < \pi(x_6),$$

$$m = 2 : \pi(x_1) < \pi(x_2) < \pi(x_3) < \pi(x_4) < \pi(x_6) < \pi(x_5).$$

One method making use of this approach is the partial order continual reassessment method, originally proposed by Wages, Conaway, and O'Quigley (2011).

The CRM for partial orders is based on using a class of working models that correspond to possible orderings of the toxicity probabilities for the combinations. Specifically, suppose there are  $M$  possible orderings being considered. For a particular ordering  $m$ , we model the true probability of toxicity at the combination  $x_i$  via a set of one-parameter working models  $F_m(x_i, \beta_m)$ , e.g., the power model

$$F_m(x_i, \beta_m) = x_{mi}^{\exp(\beta_m)}, \quad (3.7)$$

where the dose levels  $x_{mi}$  correspond to the skeleton under the working model  $m$ . The use of other single-parameter working models common to the CRM class, such as a hyperbolic tangent function or a one-parameter logistic model, have been explored, and it was found that there is little difference in the operating characteristics among the various model choices.

## Bayesian framework

Much like the CRM, the original formulation of the POCRM employed a Bayesian approach. In this setting, we begin by assigning a prior distribution,  $g_m(\beta_m)$ , for the parameter  $\beta_m$  of each model. We let the plausibility of each ordering under consideration be described by a set of prior probabilities

$$\tau = \{\tau(1), \dots, \tau(M)\},$$

where  $\tau(m) \geq 0$  and  $\tau(1) + \dots + \tau(M) = 1$ .

Suppose that, at a certain point in the trial,  $y_i$  patients have experienced DLT among the  $n_i$  patients treated on the combination  $x_i$ . Using the data  $\mathcal{D} = \{(y_i, n_i); i = 1, \dots, k\}$ , the likelihood under the  $m$ th ordering is given by

$$\mathcal{L}_m(\mathcal{D} | \beta_m) \propto \prod_{i=1}^k \{F_m(x_i, \beta_m)\}^{y_i} \{1 - F_m(x_i, \beta_m)\}^{n_i - y_i}. \quad (3.8)$$

The posterior density for  $\beta$  is then given by

$$f_m(\beta_m | \mathcal{D}) = \frac{\mathcal{L}_m(\mathcal{D} | \beta_m) g_m(\beta_m)}{\int \mathcal{L}_m(\mathcal{D} | \beta_m) g_m(\beta_m) d\beta_m}.$$

This information can be used in establishing the posterior probabilities of the models given the data as

$$P(m | \mathcal{D}) = \frac{\tau(m) \int \mathcal{L}_m(\mathcal{D} | \beta_m) g_m(\beta_m) d\beta_m}{\sum_{m=1}^M \tau(m) \int \mathcal{L}_m(\mathcal{D} | \beta_m) g_m(\beta_m) d\beta_m}.$$

The POCRM proposes to choose a single ordering (model),  $h$ , with the largest posterior model probability such that

$$h = \arg \max_m P(m | \mathcal{D})$$

and apply the Bayesian form of the CRM. Given the ordering  $h$  and the working model  $F_h(x_i, \beta_h)$ , we generate DLT probabilities at each combination so that

$$\tilde{\pi}(x_i) = F_h(x_i, \tilde{\beta}_h),$$

where

$$\tilde{\beta}_h = \int \beta_h f_h(\beta_h | \mathcal{D}) d\beta_h.$$

The combination assigned to the next patient in the trial is that which has an estimated DLT probability closest to the target DLT probability.

## POCRM in Case study 2

Returning to Case study 2, a dose-escalation trial was designed to determine the MTD combination of the treatments labeled  $x_1, \dots, x_5$  (see Table 3.3). A prior specification for the POCRM is to choose a subset of possible dose-toxicity orders based on the partial ordering of the combinations. In this example, the possibilities consist of the two orderings above. For each ordering, we need to specify a set of

skeleton values  $x_{mi}$ . We can simply specify a set of reasonably spaced values or use the algorithm of Lee and Cheung (2009) to generate values. These values are then adjusted to correspond to each of the possible orderings. For instance, for  $m = 1$  defined above, we specify  $x_{1i}$  as

$$m = 1 : 0.05 \quad 0.15 \quad 0.25 \quad 0.35 \quad \mathbf{0.45} \quad \mathbf{0.55},$$

and, for  $m = 2$  defined above, we specify  $x_{2i}$  as

$$m = 2 : 0.05 \quad 0.15 \quad 0.25 \quad 0.35 \quad \mathbf{0.55} \quad \mathbf{0.45}.$$

We assume, a priori, that each of the two orderings is equally likely and let  $\tau(1) = \tau(2) = 0.5$ . For each ordering, the prior distribution for  $\beta$  is assumed to be  $g_m(\beta_m) = N(0, 1.34)$ . DLT probabilities are modeled via the power models defined in Equation 3.7.

The resulting POCRM-based algorithm is defined as follows:

1. The first patient is administered the “lowest” combination  $x_1$ . At any point in the trial, based on the accumulated data from patients included thus far in the trial, the estimated toxicity probabilities  $\tilde{\pi}(x_i)$  are obtained for all combinations being tested, based on the procedure described above involving Bayesian model choice.
2. The next entered patient is then allocated to the dose combination with estimated toxicity probability closest to the target rate so that  $|\tilde{\pi}(x_i) - \theta|$  is minimized, with the following restriction in place: The trial is not allowed to “skip” over an untried combination when *escalating*. If all combinations have been tried, the study is allowed to move to whichever combination is recommended by the method.
3. The MTD combination is defined as the combination that minimizes  $|\tilde{\pi}(x_i) - \theta|$  after the total sample size of  $n$  patients.

In the simulation of DLT outcomes in a trial, the tolerance of each patient can be considered a uniformly distributed random variable on the interval  $[0, 1]$ . This value is termed *patient's latent toxicity tolerance* and is denoted by  $u_j$  for the  $j$ th entered patient (O'Quigley, Paoletti, and Maccario, 2002). At the combination  $(x_i)$  assigned to patient  $j$ , if the tolerance is less than or equal to its true DLT probability (i.e.,  $u_j \leq \pi(x_i)$ ), then patient  $j$  has a DLT. Otherwise the patient has a non-DLT outcome. Of course, in a real trial, it is impossible to observe a patient's latent tolerance. But it is a useful tool to be used in simulations and can be used to evaluate the operating characteristics of a dose-finding method within a single trial.

In conducting this exercise, we generated the latent tolerance sequence in Table 3.4 for  $n = 30$  patients using Program 3.6.

### PROGRAM 3.6 Latent outcome generation for 30 patients

```

data latent;
do i = 1 to 30;
  u = rand("Uniform");
  output;
end;
run;

proc print data=latent;
run;

```

<b>Output from Program 3.6</b>	<b>Obs</b>	<b>i</b>	<b>u</b>
	1	1	0.978
	2	2	0.595
	3	3	0.340
	4	4	0.055
	5	5	0.043
	6	6	0.869
	7	7	0.478
	8	8	0.649
	9	9	0.873
	10	10	0.539
	11	11	0.738
	12	12	0.129
	13	13	0.495
	14	14	0.690
	15	15	0.413
	16	16	0.787
	17	17	0.285
	18	18	0.847
	19	19	0.905
	20	20	0.963
	21	21	0.779
	22	22	0.223
	23	23	0.248
	24	24	0.862
	25	25	0.534
	26	26	0.286
	27	27	0.440
	28	28	0.237
	29	29	0.803
	30	30	0.321

The allocation algorithm is illustrated using the latent outcomes above and assessing whether each patient has a DLT using the true DLT probabilities

$$\{\pi(x_1), \dots, \pi(x_6)\} = \{0.02, 0.05, 0.08, 0.25, 0.45, 0.55\}$$

with the target rate  $\theta = 0.25$ . The method begins at the lowest combination so that Patient 1 receives  $x_1$ . Because the tolerance  $u_1 = 0.978$ , this patient does not have a DLT, since  $u_1 > 0.02$ . Escalating in cohorts of size 1, the POCRM then recommends that Patient 2 should receive  $x_2$ . The latent tolerance  $u_2 = 0.5949$  is greater than  $\pi(x_2) = 0.05$ , resulting in a non-DLT outcome. The first DLT occurs for each method at the fourth entered patient, based on a latent tolerance of  $u_4 = 0.055$ , which is less than the true DLT probability for the combination recommended to this patient.

At this point in the trial, with a limited amount of data, the estimated ordering is  $h = 1$ , and DLT probability estimates for each combination are

$$\{\tilde{\pi}(x_1), \tilde{\pi}(x_2), \tilde{\pi}(x_3), \tilde{\pi}(x_4), \tilde{\pi}(x_5), \tilde{\pi}(x_6)\} = \{0.11, 0.21, 0.30, 0.38, 0.47, 0.56\}$$

from which  $x_2$  is recommended as the combination to be administered to the next patient (i.e., Patient 5) since  $\tilde{\pi}(x_2) = 0.21$  is closest to the target value of  $\theta = 0.25$ . The process of estimation/allocation continues until, after  $n = 30$  patients, the POCRM recommends  $x_4$  as the MTD combination, which has a true DLT probability of  $\pi(x_4) = 0.25$ .

This simulated trial is shown in Table 3.4. The implementation is carried out by entering a data set and then calling the %POCRM macro. This macro has the following parameters:

**TABLE 3.4 Simulated sequential trial illustrating each approach using latent toxicity tolerance of 30 patients**

$j$	$u_j$	$x_i$	$\pi(x_i)$	$y_i$	$j$	$u_j$	$x_i$	$\pi(x_i)$	$y_i$
1	0.978	$x_1$	0.02	0	16	0.787	$x_4$	0.25	0
2	0.595	$x_2$	0.05	0	17	0.285	$x_4$	0.25	0
3	0.340	$x_3$	0.08	0	18	0.847	$x_4$	0.25	0
4	0.055	$x_4$	0.25	1	19	0.905	$x_4$	0.25	0
5	0.043	$x_2$	0.05	1	20	0.963	$x_4$	0.25	0
6	0.869	$x_1$	0.02	0	21	0.779	$x_4$	0.25	0
7	0.478	$x_2$	0.05	0	22	0.223	$x_5$	0.45	1
8	0.649	$x_2$	0.05	0	23	0.248	$x_4$	0.25	1
9	0.873	$x_2$	0.05	0	24	0.862	$x_4$	0.25	0
10	0.539	$x_2$	0.05	0	25	0.534	$x_4$	0.25	0
11	0.738	$x_3$	0.08	0	26	0.286	$x_4$	0.25	0
12	0.129	$x_3$	0.08	0	27	0.440	$x_4$	0.25	0
13	0.495	$x_3$	0.08	0	28	0.237	$x_4$	0.25	1
14	0.690	$x_3$	0.08	0	29	0.803	$x_4$	0.25	0
15	0.413	$x_3$	0.08	0	30	0.321	$x_4$	0.25	0

- **nsim** is the number of simulation runs in the simulation mode (default value is 1,000). This parameter is ignored if **realdata=TRUE**.
- **seed** is the seed for the simulation runs.
- **ncohort** is the number of patient cohort inclusions for each simulated trial.
- **startdose** is the starting dose in the simulation mode (default value is 1) and the current dose in the data analysis mode.
- **cs** is the safety threshold used to determine whether the trial should terminate for safety.
- **nstop** specifies the number of patients needed to be accrued on combination to stop the trial.
- **csize** is the size of each patient cohort (default value is 1).
- **target** is the targeted toxicity (default value is 0.25).
- **p1** is the true toxicity rate for each combination used to generate data.
- **pskel** is the skeleton of the working models corresponding to the possible orderings for the DLT probabilities.
- **realdata** determines the macro's mode (data analysis mode or simulation mode). The macro runs once and analyzes the real clinical data in the **trialdata** data set if **realdata=TRUE**. Note that the simulation settings are ignored in this case.

Application of the %POCRM macro is illustrated in Program 3.7.

### PROGRAM 3.7 POCRM implementation

```
* y is number of DLTs and n is number of patients;
data fdata;
input y n;
cards;
0 1
0 1
0 1
1 1
0 0
0 0
;
run;
```

```
*Change the start dose to the current combination after each inclusion;
*If implementation, realdata=TRUE. If simulation, realdata=FALSE;
%POCRM(nsim=1000, seed=580, ncohort=30, startdose=4, cs=0.90, nstop=61,
csize=1, target=0.25,
p1=%bquote({0.02 0.05 0.08 0.25 0.45 0.55}),
pskel=%bquote({0.05 0.15 0.25 0.35 0.45 0.55,
0.05 0.15 0.25 0.35 0.55 0.45}),
realdata=TRUE, trialdata=fdata);
```

---

**Output from  
Program 3.7**

```
mtox_sel
1

ptox_hat
0.1074933 0.2082244 0.296629 0.3826772 0.4692604 0.5577856

pstop
0

pry
0 0 0 1 0 0

prn
1 1 1 1 0 0

pselect
0 1 0 0 0 0
```

---

For the example data included in the **fdata** data set, Output 3.7 provides the estimated ordering (**mtox sel**); estimated DLT probabilities for each combination (**ptox hat**); whether the trial should be terminated (**pstop**); observed number of DLTs at each combination (**pry**); number of patients treated at each combination (**prn**); and recommended combination based on the accumulated data (**pselect**).

Program 3.8 illustrates the POCRM algorithm by simulating ten trials:

**PROGRAM 3.8 POCRM simulation**

```
%POCRM(nsim=10, seed=580, ncohort=30, startdose=1, cs=0.90, nstop=61,
csize=1, target=0.25,
p1=%bquote({0.02 0.05 0.08 0.25 0.45 0.55}),
pskel=%bquote({0.05 0.15 0.25 0.35 0.45 0.55,
0.05 0.15 0.25 0.35 0.55 0.45}),
realdata=FALSE);
```

---

**Output from  
Program 3.8**

```
pstop
0

pry
0 0.3 0.5 3 1.8 1.1

prn
1.6 2.4 7.8 11.8 4.4 2

pselect
0 0 0.2 0.6 0.1 0.1
```

---

Based on ten simulated trials, Output 3.8 provides the percentage of trials that stopped early (`pstop`); the average number of DLTs observed at each combination (`pry`); average number of patients treated at each combination (`prn`); and percentage of trials that each combination was recommended as the MTD (`pselect`).

## 3.5 Summary

---

This chapter provided an overview of dose-finding designs in early-stage dose-escalation trials with emphasis on oncology trials. It begins with a high-level summary of more straightforward approaches to defining dose-assignment methods (rule-based methods). This includes basic algorithms such as the 3+3 algorithm as well as more flexible algorithms such as group up-and-down design and mTPI method. This chapter also provided a detailed description of model-based dose-escalation methods based on the continual reassessment method (CRM). Applications of the CRM in dose-escalation trials with a single agent and with drug combinations were reviewed. The CRM serves an alternative method to the traditional rule-based methods such as the basic 3+3 algorithm. It has been extensively studied by many researchers and proven to be more versatile and efficient in determining the MTD. This class of model-based methods uses cumulative data and a flexible cohort size that allow for better decision making. We gave advice on how to implement the standard CRM algorithm and its extension to drug-combination trials (POCRM) in SAS. We hope that availability of SAS code will encourage the broader use of these methods in practice.

## 3.6 References

---

- Babb, J., Rogatko, A., Zacks, S. (1998). Cancer Phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine* 17, 1103-1120.
- Cheung, Y. K., Chappell, R. (2000). Sequential designs for Phase I clinical trials with late-onset toxicities. *Biometrics* 56, 1177-1182.
- Cheung, Y.K., Chappell, R. (2002). A simple technique to evaluate model sensitivity in the continual reassessment method. *Biometrics* 58, 671-674.
- Cheung, Y.K. (2013). Sample size formulae for the Bayesian continual reassessment method. *Clinical Trials* 10, 852-861.
- Cheung, Y.K. (2011). *Dose finding by the Continual Reassessment Method*. New York: Chapman and Hall/CRC Biostatistics Series.
- Chevret, S. (1993). The continual reassessment method in cancer Phase I clinical trials: a simulation study. *Statistics in Medicine* 12, 1093-1108.
- Dy, G.K, Bruzek, L.M., Croghan, G.A., et al. (2005). A Phase I trial of the novel farnesyl protein transferase inhibitor, BMS-214662 in combination with paclitaxel and carboplatin in patients with advanced cancer. *Clinical Cancer Research* 11, 1877-1883.
- Faries, D. (1984). Practical modifications of the continual reassessment method for phase I cancer trials. *Journal of Biopharmaceutical Statistics* 4, 147-164.
- Gatsonis, C., Greenhouse, J.B. (1992). Bayesian methods for phase I clinical trials. *Statistics in Medicine* 11, 1377-1389.

- Goodman, S.N., Zahurak, M.L., Piantadosi, S. (1995). Some practical improvements in the continual reassessment method for phase I studies. *Statistics in Medicine* 14, 1149-1161.
- Guo, W., Wang, S.J., Yang, S., Lin, S., Ji, Y. (2017). A Bayesian interval dose-finding design addressing Ockham's razor: mTPI-2. In press.
- Iasonos, A., O'Quigley, J. (2012). Interplay of priors and skeletons in two-stage continual reassessment method. *Statistics in Medicine* 31, 4321-4336.
- Iasonos, A., O'Quigley, J. (2014). Adaptive dose-finding studies: a review of model-guided phase I clinical trials. *Journal of Clinical Oncology* 32, 2505-2511.
- Iasonos, A., Wages, N.A., Conaway, M.R., Cheung, Y.K., Yuan, Y., O'Quigley, J. (2016). Dimension of model parameter space and operating characteristics in adaptive dose-finding studies. *Statistics in Medicine* 35, 3760-3775.
- Ianova, A. (2006). Escalation, up-and-down and A+B designs for dose-finding trials. *Statistics in Medicine* 25, 3668-3678.
- Ianova, I., Flournoy, N., Chung, Y. (2007). Cumulative cohort design for dose-finding. *Journal of Statistical Planning and Inference* 137, 2316-2327.
- Ianova, A., Kim, S. H. (2009). Dose finding for continuous and ordinal outcomes with a monotone objective function: a unified approach. *Biometrics* 65, 307-315.
- Ianova, A., Wang, A., Foster, M. (2016). The rapid enrollment design for Phase I clinical trials. *Statistics in Medicine* 35, 2516-2524.
- Ji, Y., Liu, P., Li, Y., Bekele, B.N. (2010). A modified toxicity probability interval method for dose-finding trials. *Clinical Trials* 7, 653-663.
- Ji, Y., Wang, S. (2013). Modified toxicity probability interval design: a safer and more reliable method than the 3+3 design for practical phase I trials. *Journal of Clinical Oncology* 31, 1785-1791.
- Lee, S.M., Cheung, Y.K. (2011). Calibration of prior variance in the Bayesian continual reassessment method. *Statistics in Medicine* 30, 2081-2089.
- Lee, S.M., Cheung, Y.K. (2009). Model calibration in the continual reassessment method. *Clinical Trials* 6, 227-238.
- Neuenschwander, B., Branson, M., Gsponer, T. (2008). Critical aspects of the Bayesian approach to Phase I cancer trials. *Statistics in Medicine* 27, 2420-2439.
- O'Quigley, J., Pepe, M., Fisher, L. (1990). Continual reassessment method: A practical design for Phase I clinical trials in cancer. *Biometrics* 46, 33-48.
- O'Quigley, J., Shen, L.Z. (1996). Continual reassessment method: A likelihood approach. *Biometrics* 52, 673-684.
- O'Quigley, J., Paoletti, X., Maccario, J. (2002). Non-parametric optimal design in dose finding studies. *Biostatistics* 3, 51-56.
- O'Quigley, J., Reiner, E.A. (1998). Stopping rule for the continual reassessment method. *Biometrika* 85, 741-748.
- Piantadosi, S., Fisher, J.D., Grossman, S.A. (1998). Practical implementation of a modified continual reassessment method for dose-finding trials. *Cancer Chemotherapy and Pharmacology* 41, 429-436.
- Rosenberger, W., Haines, L. (2002). Competing designs for phase I clinical trials: a review. *Statistics in Medicine* 21, 2757-2770.
- Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* 45, 795-798.
- Sverdlov, O., Wong, W.K., Ryeznik, Y. (2014). Adaptive clinical trial designs for Phase I cancer studies. *Statistics Surveys* 8, 2-44.

- Wages, N.A., Conaway, M.R., O'Quigley, J. (2011). Continual reassessment method for partial ordering. *Biometrics* 67, 1555-1563.
- Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society B* 25, 1-48.
- Yin, G., Yuan, Y. (2009). Bayesian model averaging continual reassessment method in Phase I clinical trials. *Journal of the American Statistical Association* 104, 954-968.
- Zink, R. Menon, S. (2016). *Clinical Trials Using SAS: Classical, Adaptive and Bayesian Methods*. SAS Press: Cary, NC.



# Chapter 4

## Dose-finding Methods

Srinand Nandakumar (Pfizer)

Alex Dmitrienko (Mediana)

Ilya Lipkovich (QuintilesIMS)

4.1	Introduction	127
4.2	Case studies	128
4.3	Dose-response assessment and dose-finding methods	132
4.4	Dose finding in Case study 1	145
4.5	Dose finding in Case study 2	160
4.6	References	176

The chapter introduces a family of statistical methods used in Phase II trials to study the relationship between the dose of an experimental treatment and clinical response. This includes methods aimed at testing the dose-response trend; estimating the underlying dose-response function; and identifying a dose or set of doses to be examined in the subsequent confirmatory trials. Powerful methods for detecting dose-response signals use one or more contrasts that evaluate the evidence of treatment benefit across the trial arms. These methods emphasize hypothesis testing with appropriate multiplicity adjustments. But they can be extended to hybrid methods that combine dose-response testing and dose-response modeling to provide a comprehensive approach to dose-response analysis (MCP-Mod procedure). The popular dose-response testing and dose-finding strategies are illustrated using Phase II trials with normally distributed and binary endpoints. Multiple practical issues arising in dose-response modeling, including adjustments for important covariates and handling of missing observations, are discussed in the chapter.

### 4.1 Introduction

---

Confirmatory Phase III trials tend to be quite expensive due to a large number of enrolled patients. It was explained in Chapter 3 that identifying an appropriate dose range for these late-stage clinical trials is one of the key goals of early-phase trials. An appropriate dose can be described as the dose with an acceptable toxicity profile that is also associated with a desirable efficacy profile. While Chapter 3 focused on dose-finding methods used in Phase I trials with dose-escalation designs, this chapter provides a survey of statistical methods used in *dose-response analysis* in Phase II trials with a parallel-group design. These trials test several doses of an experimental treatment versus a control (most commonly, a placebo) to characterize the relationship between the dose and clinical response.

The importance of correctly identifying a range of effective and tolerable doses at early stages of a development program cannot be overstated. It is surprisingly common to adjust the dose range determined in Phase III trials due to safety issues discovered in post-marketing data (Cross et al., 2002). Therefore it is critical to perform a comprehensive evaluation of the dose-response relationship in dose-finding trials. This evaluation includes several components such the detection of dose-related trends and determination of the functional form of the dose-response curve. These exercises ultimately support dose-finding strategies aimed at the identification of optimal doses--e.g., the *minimum effective dose* (MED).

The general topic of dose-response analysis has received much attention in the literature, including numerous clinical trial publications and even recent regulatory documents. For example, the European Medicines Agency released a qualification statement on a statistical methodology for model-based dose finding (EMA, 2014). A general overview of statistical considerations in dose-finding trials is provided by Ruberg (1995a, 1995b) and Chuang-Stein and Agresti (1997). A detailed discussion of various aspects of dose-response analysis can be found in Ting (2006). Statistical methods used in dose-finding trials with emphasis on SAS implementation are presented in Dmitrienko et al. (2007) and Menon and Zink (2016).

This chapter provides an overview of methods for assessing dose-response and performing dose finding commonly used in parallel-group Phase II clinical trials. Examples from dose-finding clinical trials defined in Section 4.2 are used to motivate and illustrate the methods introduced in the chapter. Section 4.3 defines a general hypothesis testing framework based on pairwise tests and dose-response contrasts. It also describes a family of algorithms based on the Multiple Comparison-Modeling (MCP-Mod) procedure that combines contrast-based tests with dose-response modeling techniques. The statistical procedures introduced in this chapter are illustrated in Sections 4.4 and 4.5.

The SAS code and data sets included in this chapter are available on the book's website at <http://support.sas.com/publishing/authors/dmitrienko.html>.

## 4.2 Case studies

---

### 4.2.1 Case study 1 (Schizophrenia clinical trial with a continuous endpoint)

Consider a Phase II trial in patients with schizophrenia. A balanced design was used in the trial with the total sample size of 240 patients. The patients were randomly allocated to four trials arms:

- Placebo
- Low dose (40 mg)
- Medium dose (80 mg)
- High dose (120 mg)

The primary endpoint was the change from baseline in the Positive and Negative Syndrome Scale (PANSS) total score to Week 4. This endpoint followed a normal distribution, and a lower value at the Week 4 visit indicated beneficial effect. The one-sided Type I error rate in this clinical trial was set to  $\alpha = 0.025$ .

The data from the schizophrenia clinical trial are included in the `cs1` data set. There are 240 rows in the data set (one record per patient) with the following variables:

- Subjid: Patient ID.
- Dose: Dose of the experimental treatment (0, 40, 80, or 120).

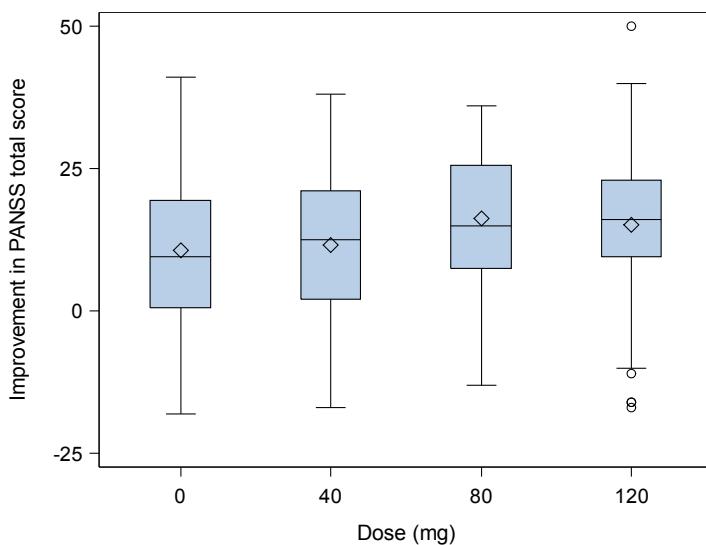
- Improvement: Reduction from baseline to Week 4 in the PANSS total score. Note that the reduction from baseline is a positive quantity, whereas the change from baseline is a negative quantity in this trial.
- Baseline: PANSS total score at baseline.
- Gender: Patient's gender.

The last two variables were included in the data set because they were likely to influence the primary endpoint, e.g., the PANSS total score at baseline was expected to be strongly correlated with the change from baseline to Week 4.

It is also worth mentioning that there were lots of missing observations in the trial. The early dropout rate was over 20% in all trial arms. For the sake of simplicity, the missing values of the primary endpoint were imputed using the basic LOCF (last observation carried forward) method. Advanced methods for handling missing data will be discussed in Case study 2.

A graphical summary of the improvement from baseline in the PANSS total score in the four arms is shown in Figure 4.1. The pattern of the mean improvement in this score across the trial arms suggests that the dose-response relationship in the schizophrenia trial might not be monotone. The mean treatment difference achieved its maximum at the 80-mg dose and a reduced effect was observed at the 120-mg dose. Dose-response functions of this kind are often referred to as umbrella functions. It is important to remember that the true dose-response relationship might in fact be linear, and the observed pattern might be due to a high level of variability in this small Phase II trial. The process of modeling dose-response relationships will be illustrated in this chapter and will help characterize the underlying dose-response in this dose-finding trial.

**Figure 4.1**  
Case study 1



*Improvement from baseline in the PANSS total score in the four trial arms.  
The mean improvement in each trial arm is represented by a diamond.*

We will use this case study to illustrate the commonly used approaches to assessing the overall evidence of a beneficial treatment effect (dose-response tests) as well as identifying the best dose-response model and target dose or doses to be examined in subsequent Phase III trials. Dose-finding methods based on simple ANOVA models with a single independent variable (dose) will be discussed in Section 4.4.1, and ANCOVA models incorporating the two predefined covariates (PANSS total score at baseline and gender) will be examined in Section 4.4.2.

## 4.2.2 Case study 2 (Urticaria clinical trial with a binary endpoint)

This case study is based on a Phase II dose-finding trial for the treatment of urticaria (hives). Patients with urticaria were randomly assigned to placebo or one of the four doses of the active treatment:

- Placebo
- Dose 1 (0.5 mg)
- Dose 2 (0.75 mg)
- Dose 3 (0.9 mg)
- Dose 4 (1 mg)

A balanced design was used with 30 patients per trial arm. There were three post-baseline visits in the trial (the visits are labeled Visit 1, Visit 2, and Visit 3). The primary analysis in the trial was based on the Hives Severity (HS) score, which was defined as the number of hives recorded by each patient in their eDiary. Based on the HS scores, the patients were categorized as responders or non-responders at each post-baseline visit.

The data from this clinical trial are included in the `cs2` data set. There are 450 observations in the data set (one record per patient and per post-baseline visit). The six variables in the data set are defined as follows:

- Subjid: Patient ID.
- Dose: Dose of the experimental treatment (0, 0.5, 0.75, 0.9, 1).
- Visit: Post-baseline visit (V1, V2, V3).
- Time: Numeric value of the post-baseline visit (1, 2, 3).
- Response: Outcome based on the HS score (0, No response; 1, Response; ., Missing).
- Gender: Patient's gender.

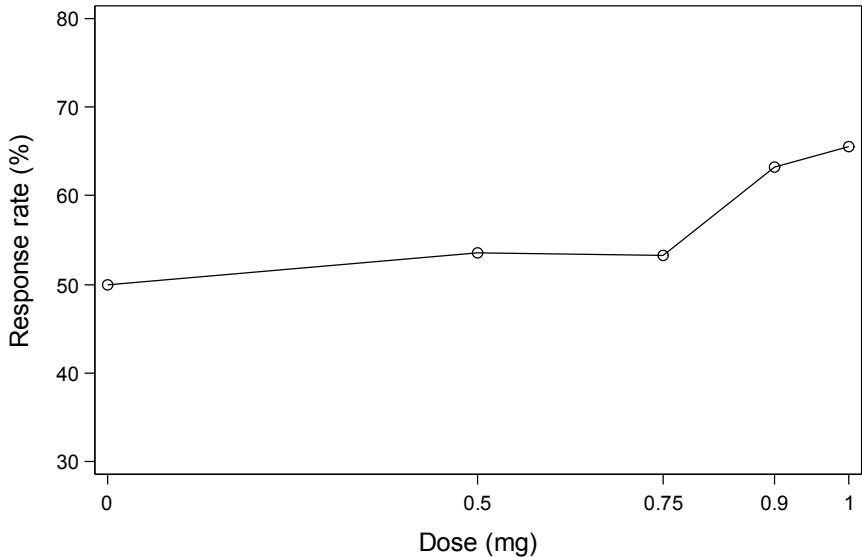
It is important to note that some patients missed one of the post-baseline visits, and hence their responder status was not ascertained at those visits. Table 4.1 presents a summary of responses and missed visits in the urticaria trial. It follows from the table that one patient missed Visit 1, two patients missed Visit 2, and five patients missed the last visit.

**TABLE 4.1 Summary of outcomes by trial arm and visit in Case study 2**

Trial arm	Outcome	Visit 1	Visit 2	Visit 3
Placebo	Missing	1	0	2
	No response	12	18	14
	Response	17	12	14
0.5 mg	Missing	0	0	2
	No response	12	7	13
	Response	18	23	15
0.75 mg	No response	13	15	14
	Response	17	15	16
0.9 mg	Missing	0	2	0
	No response	10	7	11
	Response	20	21	19
1 mg	Missing	0	0	1
	No response	9	11	10
	Response	21	19	19

A summary of response rates at the end of the trial (i.e., at Visit 3) is provided in Figure 4.2. This summary was created using a simplistic approach to handling missing observations, i.e., the five patients who missed the last visit were removed from the analysis. As shown in the figure, the response rate based on the HS score increased monotonically with dose. The treatment effect was quite small at 0.5 mg and 0.75 mg but improved considerably at the two highest doses. The overall dose-response followed a logistic or sigmoid curve with a steep increase between 0.75 mg and 0.9 mg.

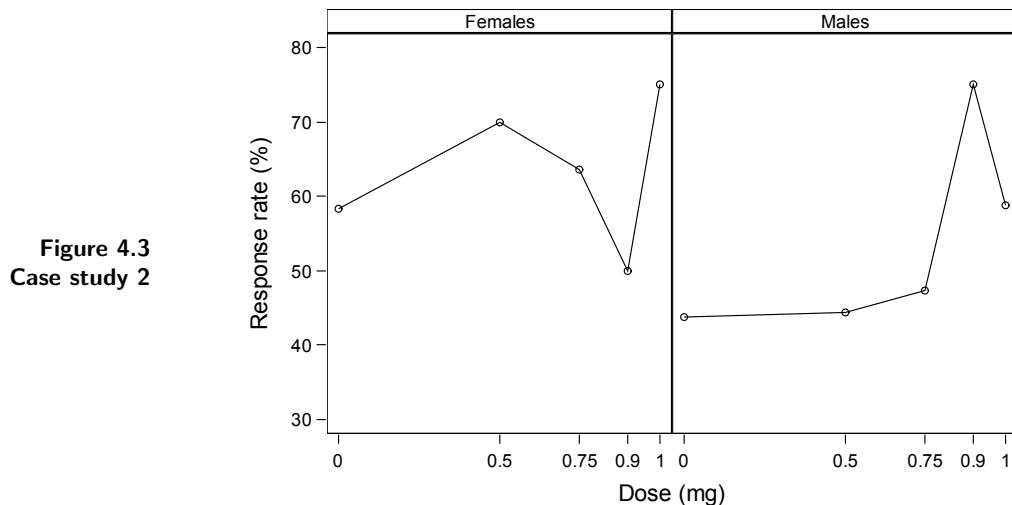
**Figure 4.2**  
Case study 2



*Response rate at Visit 3 as a function of the dose.*

An important feature of this data set is that there were more males than females enrolled in the trial (only 40% of the patients were females). The trial's sponsor suspected that gender may have predictive properties in this setting, i.e., it might be used to predict treatment response. Figure 4.3 shows the response rates by gender. It is demonstrated in this figure that the dose-response relationship was fairly monotone with a drop at the highest dose in the subset of male patients. By contrast, in the female subgroup, the dose-response had an unusual zigzag shape. Again, given that the number of females in each trial arm was quite small, it might be premature to conclude that the experimental treatment was effective mostly in males. However, if the sponsor had reasons to believe that the treatment effect would be influenced by gender, this factor clearly needs to be incorporated into the analysis model.

This case study will be used to illustrate dose-finding methods in trials with repeated measures and missing observations. Section 4.5.1 will present a dose-finding strategy aimed at finding an appropriate target dose based on a repeated-measures model for the binary responses. This strategy incorporates an adjustment for the important baseline factor (patient's gender). This strategy will be applied without accounting for the missed visits. Further, Section 4.5.2 will discuss dose-finding approaches in the presence of missing data.



**Figure 4.3**  
**Case study 2**

*Response rate at Visit 3 by gender as a function of the dose.*

## 4.3 Dose-response assessment and dose-finding methods

---

The design of parallel-group Phase II trials is similar to the design of trials employed in subsequent Phase III trials (e.g., multiple doses or regimens of a novel treatment are compared to a placebo). However, there are important differences in the overall objectives of Phase II and Phase III trials that impact the analysis of dose-placebo comparisons and dose-related trends. The main goal of examining several doses in Phase II trials is to support dose finding and select the most promising doses to be tested in the next phase. For this reason, global tests (e.g., contrast-based tests), are broadly used in Phase II trials to evaluate the overall effect of the experimental treatment across the dose range. Phase II trials might not be powered to examine the treatment effect at each dose compared to placebo, and, as a result, the individual dose-placebo tests may be non-informative. When multiple doses are studied in Phase III trials, the analysis is performed to support effectiveness claims for the individual doses, i.e., each dose is declared effective or ineffective. From this perspective, global trend tests become less relevant, and the focus shifts to examining the treatment effect at each dose based on multiple dose-placebo comparisons.

In Section 4.3.1, we will briefly discuss pairwise tests in dose-finding trials. This approach will be compared with contrast-based tests in Sections 4.3.2 and 4.3.3. Two different approaches will be described in these two sections. The first one relies on basic tests with a single set of contrast coefficients, and the second one uses inferences based on a joint analysis of several contrasts. The discussion of multi-contrast tests will provide a foundation for the more advanced dose-finding algorithms based on the MCP-Mod procedure that will be presented in Section 4.3.4 and later illustrated in Sections 4.4 and 4.5.

### 4.3.1 Dose-placebo tests

Our discussion of statistical tests aimed at detecting dose-related trends in Phase II trials begins with a simplistic approach that relies on pairwise comparisons. The tests considered in this section, as well as more powerful contrast-based tests presented in Sections 4.3.2 and 4.3.3, will be defined using a dose-finding trial with  $m + 1$

arms ( $m$  doses of an experimental treatment versus placebo) and a balanced design with  $n$  patients per arm. The primary endpoint will be assumed to be normally distributed, and a larger value of the endpoint is associated with improvement. A simple ANOVA model will be considered, i.e.,

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where  $y_{ij}$  is the outcome of the  $j$ th patient in the  $i$ th arm,  $i = 0, \dots, m$  and  $j = 1, \dots, n$ . The error term  $\varepsilon_{ij}$  is normally distributed with mean 0 and variance  $\sigma^2$ .

Dose-placebo tests can be carried out in this Phase II trial to test the following null hypotheses of no treatment effect:

$$H_i : \mu_0 = \mu_i, \quad i = 1, \dots, m,$$

versus alternative hypotheses of a positive treatment effect in the  $m$  dosing arms.

The schizophrenia clinical trial from Case study 1 introduced in Section 4.2 can be used to illustrate this setting with  $m = 3$  dosing arms. The dose-placebo comparisons are easy to perform in this trial using a number of procedures in SAS/STAT, including PROC GLM, PROC MIXED, and PROC GENMOD. Program 4.1 uses PROC MIXED to carry out the three dose-placebo tests using the improvement from baseline to Week 4 in the PANSS total score as the primary endpoint. The significance of each dose-placebo comparison is assessed using the simple ANOVA model defined above. In addition, a multiplicity adjustment derived from the Dunnett test is requested using the ADJUST statement. This adjustment results in a set of multiplicity-adjusted  $p$ -values and inferences based on the adjusted  $p$ -values control the overall Type I error rate in this trial. The Dunnett test is an example of efficient *parametric* multiple tests that account for positive correlations among the test statistics in dose-finding trials. For example, the test statistics are known to be equally correlated with the common correlation coefficient  $\rho = 0.5$  in dose-finding trials with a balanced design. The Dunnett test results in more powerful inferences compared to the basic Bonferroni test in both balanced and unbalanced designs. See Chapter 5 for a detailed discussion of multiple tests commonly used in clinical trials to protect the overall Type I error rate.

#### PROGRAM 4.1 Dose-placebo tests in Case study 1

```
proc mixed data=cs1;
  class dose;
  model improvement=dose;
  lsmeans dose/pdiff adjust=dunnett;
run;
```

The output of Program 4.1 is summarized in Table 4.2. The table lists the test statistics for the individual dose-placebo comparisons and associated two-sided  $p$ -values computed from central  $t$  distributions (raw  $p$ -values). Using a two-sided  $\alpha = 0.05$  as the threshold that defines statistical significance, it is clear that the treatment difference is significant only at 80 mg ( $p = 0.0242$ ). A trend toward statistical significance is detected at 120 mg ( $p = 0.0742$ ).

**TABLE 4.2 Output from Program 4.1**

Dose	Treatment difference	Standard error	Test statistic	<i>P</i> -value	
				Raw	Dunnett-adjusted
40 mg	0.9167	2.5279	0.36	0.7172	0.9688
80 mg	5.7333	2.5279	2.27	0.0242	0.0634
120 mg	4.5333	2.5279	1.79	0.0742	0.1801

It is important to bear in mind that the simultaneous analysis of the three dose-placebo comparisons induces multiplicity. If no multiplicity adjustment is applied, the inferences are no longer reliable due to an inflated Type I error rate. To preserve the error rate, inferences must be performed using multiplicity-adjusted  $p$ -values, e.g., the Dunnett-adjusted  $p$ -values listed in the table. To analyze the treatment effect using the adjusted  $p$ -values, each adjusted  $p$ -value needs to be compared to the same threshold as before, i.e.,  $\alpha = 0.05$ . None of the Dunnett-adjusted  $p$ -values exceeds the threshold, which implies that there is no longer evidence of a significant treatment effect after the multiplicity is accounted for.

If the schizophrenia clinical trial was a confirmatory Phase III trial, the results presented in Table 4.2 would lead to the conclusion that this trial failed to achieve its primary endpoint. However, as stressed above, dose-finding trials are typically not powered to support pairwise comparisons. A more sensible approach in this setting is to set up tests that evaluate dose-related trends in the outcome variable. This can be easily accomplished using contrast-based tests discussed in Sections 4.3.2 and 4.3.3.

### 4.3.2 Contrast-based tests

The dose-placebo tests considered in Section 4.3.1 are a special case of a broad family of contrast-based tests. The main reason that pairwise tests tend to be underpowered in Phase II trials is that they focus on individual dose-placebo comparisons and thus do not use all of the available information. General contrast-based tests gain power by pooling the evidence of treatment benefit across multiple dosing arms. These tests serve as efficient tools for detecting dose-related trends and have found numerous applications in dose-finding trials.

A dose-response contrast in a clinical trial with  $m + 1$  arms is constructed by specifying a vector of arm-specific coefficients denoted by  $c_0, \dots, c_m$ . The coefficients are selected to add up to 0, i.e.,

$$\sum_{i=0}^m c_i = 0,$$

and they are often standardized to ensure that

$$\sum_{i=0}^m c_i^2 = 1.$$

Given a set of contrast coefficient, the contrast is defined as

$$c = \sum_{i=0}^m c_i \hat{\mu}_i,$$

where  $\hat{\mu}_0, \dots, \hat{\mu}_m$  denote the sample means in the  $m + 1$  trial arms. For example, the test that evaluates the effect of the  $i$ th dose compared to placebo corresponds to the contrast with  $c_0 = -1$ ,  $c_i = 1$  and  $c_j = 0$  if  $i \neq j$ .

In the context of contrast-based tests, the null hypothesis being tested is the null hypothesis of no treatment effect across the  $m$  trial arms:

$$H : \sum_{i=0}^m c_i \mu_i = 0.$$

Contrast-based tests are set up as one-sided tests, which means that the null hypothesis is rejected in favor of the alternative defined as

$$K : \sum_{i=0}^m c_i \mu_i > 0$$

if the contrast's value is large. A rejection of the null hypothesis provides evidence of a significant dose-response relationship in the trial.

To evaluate the significance of the dose-response trend using the selected contrast, the following  $t$  statistic is computed by dividing the contrast by its standard error, i.e.,

$$t = \frac{\sum_{i=0}^m c_i \hat{\mu}_i}{\sigma \sqrt{\sum_{i=0}^m c_i^2 / n}}.$$

This test statistic follows a  $t$  distribution, e.g., in a balanced setting with  $n$  patients per arm, it is  $t$ -distributed with  $(m + 1)(n - 1)$  degrees of freedom. A significant test statistic is interpreted as a demonstration of the treatment's efficacy.

It is helpful to compare the general family of contrast-based tests in a simple ANOVA setting to the the  $F$  test. The latter is aimed at assessing the amount of evidence to reject the overall null hypothesis of no effect across the trial arms (tests of this kind are known as omnibus tests), i.e.,

$$H : \mu_0 = \mu_1 = \dots = \mu_m.$$

The overall null hypothesis is symmetrical with respect to all possible permutations of the  $m + 1$  arms, and the  $F$  test ignores the dose order information. As a direct consequence of this, contrast-based tests are more powerful than the basic  $F$  test in dose-finding trials.

Numerous types of contrasts or, in other words, specific sets of contrast coefficients have been proposed in the literature. These tapes include basic contrasts that rely on the assumption of a linear dose-response and a variety of more advanced contrasts. As a quick illustration, suppose that the dose levels are equally spaced in a dose-finding trial with  $m + 1$  arms. In this case, the linear contrast is set up by assigning integer scores from 0 to  $m$  to the individual arms in the trial and then subtracting the mean to ensure that the resulting coefficients sum to 0, i.e.,

$$c_i = i - m/2, \quad i = 0, \dots, m.$$

It is easy to verify that the coefficients of the linear contrast in a four-arm trial (e.g., the schizophrenia trial from Case study 1) are given by  $c_0 = -3$ ,  $c_1 = -1$ ,  $c_2 = -1$  and  $c_3 = 3$ .

Simple tests similar to tests derived from the linear contrast were widely used by the middle of the 20th century. See, for example, Scheffe (1959). More advanced contrasts with optimal properties started attracting attention in the 1960s. To give a quick example, Abelson and Tukey (1963) introduced the modified linear test. This contrast provides an easy-to-implement approximation to the maximum contrast (also introduced by Abelson and Tukey [1963]), that maximizes power among all tests against the worst-case configuration under the ordered alternative. To improve power, the modified linear test assigns larger coefficients to the lower and higher doses. For example, in the context of Case study 1, the coefficients of the modified linear contrast are  $c_0 = -12$ ,  $c_1 = -2$ ,  $c_2 = -2$ , and  $c_3 = 12$ . Other examples of advanced contrasts include the regular and reverse Helmert contrasts as well as contrasts based on isotonic regression (Williams, 1971, 1972).

Program 4.2 illustrates the process of carrying out contrast-based tests in Case study 1. Using PROC MIXED, as in Program 4.1, this program includes CONTRAST statements to evaluate the significance of dose-related trends in the trial using the linear and modified linear contrasts defined above. The CONTRAST statements specify the contrast coefficients for the four arms in the schizophrenia trial.

## PROGRAM 4.2 Linear and modified linear contrast tests in Case study 1

```
proc mixed data=cs1;
   ods select tests3 contrasts;
   class dose;
   model improvement=dose;
   contrast "Linear" dose -3 -1 1 3;
   contrast "Modified linear" dose -12 -2 2 12;
run;
```

Table 4.3 lists the statistics and two-sided  $p$ -values for the two trend tests produced by Program 4.2. Since the evidence of treatment effect is pooled across the four arms, the test statistics of the linear and modified linear contrasts are much greater than the statistics displayed in Table 4.2. Both two-sided  $p$ -values are significant at a 0.05 level. Based on each  $p$ -value, we can conclude that there is a significant dose-related trend in the primary endpoint in Case study 1.

**TABLE 4.3 Output from Program 4.2**

Contrast	Test statistic	$P$ -value
Linear	5.31	0.0221
Modified linear	4.34	0.0384

It is also instructive to compare the two trend tests with the  $F$  test for testing the null hypothesis of no treatment effect in the schizophrenia trial. In this case, the  $F$  statistic computed from the ANOVA model is equal to 2.4. Since the numerator and denominator degrees of freedom are equal to 3 and 236, respectively, the two-sided  $p$ -value is not significant at a 0.05 level ( $p = 0.0686$ ). As emphasized above, the  $F$  test is an example of an omnibus test and does not take into account the dose ordering that results in power loss compared to contrast-based tests.

### 4.3.3 Multi-contrast tests

As demonstrated in Section 4.3.2, trend tests based on dose-response contrasts provide improvement over pairwise tests in dose-finding trials. An important feature of contrast-based tests is that they are most powerful when the shape of the true dose-response function matches the pattern of the contrast coefficients. For example, the power of a trend test based on the linear contrast introduced above is maximized when the underlying dose-response function is also linear. If the assumption of linearity is violated, especially when the dose-response is non-monotonic, the linear contrast test will perform poorly.

To address this limitation of contrast-based tests, we can replace tests that rely on a single dose-response contrast, known as *single-contrast tests*, with dose-response tests that incorporate several contrasts. These tests are termed *multi-contrast tests* and exhibit a more robust behavior if little is known about the shape of the dose-response function.

With a multi-contrast testing approach, a set of contrasts is predefined in a dose-finding trial. The contrasts may be selected from the same family (e.g., a family of Helmert contrasts) or represent a combination of contrasts from different families. Let  $k$  denote the number of contrasts and  $t_1, \dots, t_k$  denote the contrast-specific test statistics. It can be shown that the test statistics follow a multivariate  $t$  distribution, and it is important to note that the correlation matrix of this distribution is known at the design stage. This matrix is a simple function of the contrast coefficients and sample sizes in the trial arms. A number of testing strategies can be constructed based on these statistics, e.g., the contrast statistics can be examined based on a pre-

specified testing sequence as in the fixed-sequence procedure defined in Chapter 5. Alternatively, a step-down approach with a data-driven testing sequence can be used. In this case, testing begins with the largest statistic. If it is significant, based on an appropriate critical value, the second largest test statistic is considered, etc. The resulting multiple test is, in fact, an extension of the step-down Dunnett test that will be introduced in Chapter 5. The resulting set of significant contrasts informs decisions related to dose selection. For more information on tests based on multiple contrasts in clinical trial applications, including the problem of identifying the minimum effective and maximum safe doses, see Ruberg (1989) and Tamhane and Logan (2002). SAS implementation of multi-contrast tests in dose-finding trials is discussed in Dmitrienko et al. (2007).

#### 4.3.4 Multiple Comparison-Modeling (MCP-Mod) procedure

The contrast-based approach introduced in Sections 4.3.2 and 4.3.3 serves as a foundation for an efficient dose-finding algorithm known as the MCP-Mod which combines algorithms that rely on multiple comparison procedures (MCP), e.g., multi-contrast tests, and dose-response modeling approaches that are broadly used in early-phase trials.

The key idea behind the MCP-Mod procedure is that the sponsor of a dose-finding Phase II trial needs to simultaneously account for the uncertainty around the dose-response model and protect the probability of an incorrect conclusion about the dose-response (Type I error rate). The shape of the true dose-response function is unknown at the beginning of a Phase II trial, and model uncertainty in the context of dose-finding trials was first described in Tukey et al. (1985). The issue of model uncertainty can be addressed, to a degree, within the contrast-based framework. For example, instead of focusing on a single contrast that might not be representative of the true dose-response model, the trial's sponsor can begin with a set of candidate contrasts and set the goal of demonstrating that at least one contrast-based test produces a significant result. As explained in Section 4.3.3, this approach has been successfully used to set up multi-contrast tests in Phase II trials. A similar idea was considered by Stewart and Ruberg (2000), who investigated the performance of a two-contrast procedure to build a robust tool for assessing dose-response trends.

However, within the contrast-based framework, a hypothesis testing approach is emphasized with appropriate multiplicity adjustments, but no estimation methods are supported, e.g., the functional form of the true dose-response model cannot be estimated. The dose is treated as an ordinal variable at best, rather than a continuous variable. As a direct result of this, after the global hypothesis of no treatment effect is rejected and the experimental treatment is shown to be effective, it is not clear how to approach the problem of estimating target doses.

To address this limitation of the contrast-based testing approach, estimation is embedded into a modeling framework. With this approach, the hypothesis testing framework is expanded to provide estimates of the underlying dose-response function and target doses. This can be accomplished by exploiting the relationship between dose-response models and dose-response contrasts described in Section 4.3.3. Beginning with a family of models that incorporate historical information, the sponsor of a Phase II trial can identify the set of contrasts whose coefficients have the same shape as the candidate models. Standard contrast-based inferences are then performed, including the assessment of dose-response relationships and selection of most relevant contrasts while controlling the probability of an incorrect selection via a multiplicity adjustment. After the set of clinically relevant contrasts has been identified, the sponsor can switch back to the dose-response models that correspond to the chosen contrasts and focus on the problem of dose selection.

This general approach was first used to define the MCP-Mod procedure in Bretz, Pinheiro, and Branson (2005). Since then, several book chapters and publications have incorporated this approach. The EMA statement mentioned in the Introduction identified the MCP-Mod as a qualified efficient statistical methodology for model-based design and analysis of Phase II dose finding studies under model uncertainty (EMA, 2014). For a detailed discussion of the general MCP-Mod methodology and related approaches, see Pinheiro, Bretz, and Branson (2006) and Bretz, Tamhane, and Pinheiro (2009). The original procedure was later extended to multiple settings, e.g., to dose-finding trials with a binary endpoint (Klingenberg, 2009), general parametric models (Pinheiro et al., 2013), models for non-normal outcome variables (Pinheiro, Bretz, and Bornkamp, 2014), as well as response-adaptive designs (Bornkamp et al., 2011).

## Dose finding based on the MCP-Mod procedure

To address the general goals formulated above, the MCP-Mod procedure uses the following dose-finding algorithm:

- Step 1: Define a set of candidate dose-response models.
- Step 2: Derive the optimal contrasts from the candidate models.
- Step 3: Carry out the dose-response tests based on the optimal contrasts.
- Step 4: Select the clinically relevant dose-response models.
- Step 5: Identify the target dose or doses using the best dose-response models.

We will begin with a review of statistical methods applied at each step of this algorithm. Detailed descriptions of SAS tools that are used in the dose-finding algorithm are provided later in this section.

Consider a dose-finding clinical trial with a normally distributed endpoint. Using the general set-up introduced at the beginning of this section, assume that  $m$  doses of an experimental treatment are tested versus placebo. The dose levels are denoted by  $d_0, \dots, d_m$  with  $d_0$  corresponding to the placebo arm. For simplicity, a balanced design with  $n$  patients per arm will be assumed in this trial. The setting is easily extended to clinical trials with unbalanced designs and outcome variables that do not follow a normal distribution.

### Step 1: Candidate models

We will assume at first that the outcome variable is modeled as a function of the dose and that no other covariates are included in the model. In other words, the one-way ANOVA model introduced in Section 4.3.1 will be considered, i.e.,

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where the mean effect  $\mu_i$  depends on the dose-response model  $f(\cdot)$ , dose  $d_i$ , and a vector parameter of unknown parameters, denoted by  $\boldsymbol{\theta}$ , that describes the true dose-response relationship, i.e.,

$$\mu_i = f(d_i, \boldsymbol{\theta}).$$

As before,  $y_{ij}$  denotes the outcome of the  $j$ th patient in the  $i$ th arm,  $i = 0, \dots, m$  and  $j = 1, \dots, n$ , and the error term  $\varepsilon_{ij}$  follows a normal distribution with mean 0 and variance  $\sigma^2$ . The general case of ANCOVA models with terms corresponding to baseline covariates will be discussed in Section 4.4.2.

Dose-finding will be performed using a set of  $r$  candidate dose-response models. In general, the initial selection of candidate models is driven by available historical information. The candidate models are set up using the hypothesized dose-response functions as follows:

$$f(d, \boldsymbol{\theta}) = a + bg(d, \boldsymbol{\theta}_0), \quad (4.1)$$

where  $g(\cdot)$  is the standardized form of the dose-response function with  $\boldsymbol{\theta}_0$  denoting the vector of initial values of the model parameters or guesstimates that are derived from the expected shape of the dose-response relationship. Further,  $a$  is the location parameter, and  $b$  is the scale parameter. The reason the standardized form is introduced is that the optimal contrasts to be defined in Step 2 depend only on the standardized dose-response model.

Frequently used dose-response models are listed in Table 4.4. It is easy to see that the number of model parameters is reduced when a model is standardized. For example, there are no parameters left in the standardized form of the Linear model.

**TABLE 4.4** Commonly used dose-response models

Model	$f(d, \boldsymbol{\theta})$	$g(d, \boldsymbol{\theta}_0)$
Linear	$E_0 + \delta d$	$d$
LogLinear	$E_0 + \delta \log(d + c)$	$\log(d + c)$
Emax	$E_0 + E_{\max}d/(ED_{50} + d)$	$d/(ED_{50} + d)$
Exponential	$E_0 + E_1 [\exp(d/\delta) - 1]$	$\exp(d/\delta) - 1$
Logistic	$E_0 + E_{\max}/(1 + \exp \frac{ED_{50}-d}{\delta})$	$1/(1 + \exp \frac{ED_{50}-d}{\delta})$
SigEmax	$E_0 + E_{\max}d^h/(ED_{50} + d^h)$	$d^h/(ED_{50} + d^h)$
Quadratic	$E_0 + \beta_1 d + \beta_2 d^2$	$d + \beta_2 d^2/ \beta_1 $
Beta	$E_0 + E_{\max}\beta(\alpha, \beta)(d/D)^\alpha(1 - d/D)^\beta$	$\beta(\alpha, \beta)(d/D)^\alpha(1 - d/D)^\beta$

## Step 2: Optimal contrasts

As explained earlier in this section, the MCP-Mod procedure takes advantage of an important fact that a unique dose-response contrast is associated with any dose-response model. This contrast is defined as the contrast that has the highest power, i.e., the highest probability of rejecting the associated null hypothesis of no effect, under this particular model.

To derive the optimal dose-response contrast for a model, consider the  $i$ th model from the candidate set. The standardized version of the model, denoted by  $g_i(\cdot)$ , is found, and the following set of predicted effects is computed under this standardized model:

$$u_{ij} = g_i(d_j, \boldsymbol{\theta}_0), \quad j = 0, \dots, m.$$

It follows from this formula that the predicted effects are obtained using the initial values of the model parameters ( $\boldsymbol{\theta}_0$ ).

Using a balanced design for illustration, the optimal contrast is found from the vector of predicted effects in a very straightforward way---namely, by centering and

then normalizing this vector. The coefficients of the optimal contrast for the  $i$ th model are defined as follows:

$$c_{ij} = (u_{ij} - \bar{u}_i) / \sqrt{\sum_{l=0}^m (u_{il} - \bar{u}_i)^2}, \quad j = 0, \dots, m, \text{ where } \bar{u}_i = \frac{1}{m+1} \sum_{l=0}^m u_{il}.$$

The algorithm for finding optimal contrasts in general unbalanced designs is given in Bretz, Pinheiro, and Branson (2005).

### **Step 3: Dose-response tests**

The optimal contrast is computed for each of the  $r$  models in the candidate set as explained in Step 2. The significance of the dose-response trend based on each model-specific contrast is assessed using the  $t$  statistics defined in Section 4.3.1. These tests are aimed at detecting the existence of a dose-response relationship, and the resulting test statistics are denoted by  $t_1, \dots, t_r$ .

The dose-response tests play a central role in the MCP-Mod dose-finding algorithm since they help identify the set of clinically relevant models that will then be used to formulate recommendations for dose selection for the confirmatory trials. An important component of this process is an adjustment for multiplicity, which is applied to control the probability of incorrect model selection. The multiplicity adjustment is conceptually similar to the Dunnett test used in Section 4.3.1 in the sense that it also relies on a parametric multiple testing procedure. The joint distribution of the  $r$  test statistics is taken into account to compute an adjusted critical value that defines the threshold for statistical significance.

It can be shown that, under the global null hypothesis of no treatment effect under any model, the test statistics follow a central multivariate  $t$  distribution with  $\nu$  degrees of freedom and correlation matrix  $R$ . Here  $\nu = N - (m + 1)$ , where  $N$  is the total sample size in the trial, and  $m + 1$  is the number of trial arms. Further, in a simple setting where the candidate models do not include covariates, the correlation coefficients are known at the design stage. Assuming a balanced design with  $n$  patients per arm, the correlation between the test statistics  $t_i$  and  $t_j$ ,  $i \neq j$  is given by

$$\rho_{ij} = \frac{\sum_{l=0}^m c_{il} c_{jl}}{\sqrt{\sum_{l=0}^m c_{il}^2 \sum_{l=0}^m c_{jl}^2}}.$$

In a general case of covariate-adjusted inferences, the correlation matrix of this multivariate distribution is estimated from the general parametric model using an appropriate method, including maximum likelihood or generalized estimating equations (Pinheiro et al., 2013).

To derive the multiplicity-adjusted critical value, denoted by  $q$ , let  $T_1, \dots, T_r$ , denote the random variables that have the same joint distribution as  $t_1, \dots, t_r$  under the global null hypothesis. The adjusted critical value is found from

$$P\{T_1 \geq q \text{ or } \dots \text{ or } T_r \geq q\} = \alpha.$$

Note that the term on the left side represents the probability of erroneously identifying at least one model as significant under the assumption that there are no meaningful dose-response models. This probability of incorrect selection is set to the predefined  $\alpha$ , e.g., one-sided  $\alpha = 0.025$ . Since the joint distribution of the test statistics is known, standard numerical integration routines for the evaluation of

multivariate  $t$  probabilities, e.g., the randomized quasi-Monte Carlo methods of Genz and Bretz (2002), can be used to compute the adjusted critical value.

The final set of clinically relevant dose-response models contains the models associated with a significant dose-response relationship at the adjusted level, i.e., the models with  $t_i \geq q$ ,  $i = 0, \dots, m$ . The number of models in this set is denoted by  $s$ . It needs to be stressed that the adjusted clinical value and, as a result, the model selection rule depend on the initial values of the model parameters in the candidate set.

## Step 4: Model selection

Using the set of  $s$  clinically relevant dose-response models, the next step involves identifying the best dose-response contrast. It can be defined as the contrast that corresponds to

- The maximum test statistic
- The best information criterion

The first selection rule relies only on the strength of evidence against the contrast-specific null hypothesis of no effect, and the second rule uses information on the functional form of each dose-response model. This rule can be constructed using any of the popular information criteria such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC). For a dose-response model with  $p$  parameters, the AIC and BIC are defined as follows:

$$\text{AIC} = n + n \log 2\pi + n \log(\text{RSS}/n) + 2(p + 1)$$

and

$$\text{BIC} = n + n \log 2\pi + n \log(\text{RSS}/n) + \log(n)(p + 1),$$

where RSS is the residual sum of squares. It is also worth mentioning the corrected AIC (AICc), which is generally recommended when the sample size is relatively small, which is the case in most dose-finding trials. For more information on these and related information criteria, see, for example, Claeskens and Hjorth (2008). The best model is selected as the model that minimizes the predefined information criterion. Once the best contrast has been found, the parameters of the corresponding dose-response model are estimated.

Even though it is common to focus on the single best model based on the maximum test statistic or appropriate information criterion, the trial's sponsor could also consider selecting a set of two or more promising dose-response models.

## Step 5: Target dose selection

Given a clinically relevant dose-response model chosen in Step 4 that provides the best fit to the data, we can proceed to estimating the target dose or doses of interest. The goal of selecting the minimum effective dose (MED) as well as a range of effective doses to be examined in confirmatory trials is pursued in most Phase II trials. The MED is commonly defined as the smallest dose that ensures a clinically meaningful and statistically significant improvement over placebo (Ruberg, 1995a). If  $\Delta$  denotes a prespecified threshold that determines the clinically relevant effect, the true MED is derived as the smallest dose  $d$  for which the treatment difference, i.e., the difference between  $f(d)$  and  $f(d_0)$ , is greater than  $\Delta$ . Here  $f(d)$  is the mean

effect at the dose  $d$ , and  $f(d_0)$  is the mean effect in the placebo arm under the selected model. In other words,

$$\text{MED} = \min\{d : f(d) > f(d_0) + \Delta\}.$$

This derivation relies on the true dose-response function  $f(d)$ , but it is easily extended to the case when the dose-response function is unknown to find an estimated minimum effective dose. For example, the MED is often estimated as follows

$$\widehat{\text{MED}} = \min\{d : \widehat{f}(d) > \widehat{f}(d_0) + \Delta \text{ and } L(d) > \widehat{f}(d_0)\},$$

where  $\widehat{f}(d)$  is the predicted mean response at the dose  $d$ ;  $\widehat{f}(d_0)$  is the predicted mean response in the placebo arm; and  $L(d)$  is the lower bound of the pointwise confidence interval for  $f(d)$  with an appropriate coverage probability. This means that, at the estimated MED, the predicted mean treatment difference exceeds the clinically relevant threshold, and, in addition, it is statistically significant. Further, the best dose-response model identified in Step 4 can be used to estimate the doses that correspond to a given fraction of the maximum treatment effect, e.g., the ED<sub>80</sub> dose is the dose that provides 80% of the maximum effect under the selected model.

The discussion presented above assumed that a single model is identified at the end of Step 4 to describe the dose-response relationship in the trial. If several dose-response models are selected, model averaging techniques can be applied to set up a weighted model and apply it to estimate the MED. Model averaging for MED estimation is often performed by approximating model weights using an information criterion such as the AIC or BIC, as suggested in Pinheiro et al. (2013). Suppose that  $k$  dose-response models were identified as the most promising in Step 4, and let  $p_1, \dots, p_k$  denote the prior model probabilities that quantify how well these models are believed to approximate the true dose-response function. For example, if the dose-response relationship was believed to be linear prior to the start of the dose-finding trial, the Linear model would be assigned a high prior probability. In case no clinical information is available to define the prior model probabilities, it is reasonable to assume a non-informative prior, i.e., let  $p_1 = \dots = p_k = 1/k$ .

Let  $\text{IC}_i$  denote the value of the appropriate information criterion for the  $i$ th model,  $i = 1, \dots, k$ . The weight of the  $i$ th model is then defined as follows:

$$w_i = \frac{p_i \exp(-\text{IC}_i/2)}{\sum_{j=1}^k p_j \exp(-\text{IC}_j/2)}.$$

Lastly, let  $\widehat{\text{MED}}_i$  denote the minimum effective dose computed from the  $i$ th model. The MED is estimated from the selected dose-response models as a weighted sum of the model-specific minimum effective doses, i.e.,

$$\widehat{\text{MED}} = \sum_{i=1}^k w_i \widehat{\text{MED}}_i.$$

It is worth noting that this algorithm approaches model averaging from a rather ad hoc frequentist perspective, where AIC or BIC criteria are used to approximate model weights (see Hoeting et al., 1999). The BIC was suggested in Raftery (1995) as a simple approximation to the integrated likelihood given the selected model, and Buckland et al. (1997) suggested the use of the AIC from different perspectives. See also Kass and Raftery (1995) on the relative merits of the BIC and AIC in this context.

## Implementation of MCP-Mod procedure

The SAS implementation of the MCP-Mod dose-finding algorithm will be illustrated in Section 4.4 using Case study 1 with a continuous endpoint and in Section 4.5 using Case study 2 with a binary endpoint. The implementation relies on the following steps and macros written by the authors.

Beginning with Step 1 of the dose-finding algorithm, we need to specify the list of candidate dose-response models. In general, care must be taken in selecting similar fits such as the logistic and SigEmax models as their correlation approaches 1. As detailed in Section 4.4.3, it is justified to include only one of these models in the candidate set as they predict nearly identical dose-response profiles. Further, the trial's design needs to be specified, including the dose levels and sample sizes in each trial arm. This is accomplished by defining the `models` and `dose` data sets as shown below. By default, the dosage in the placebo arm dose is assumed to be 0.

```

data models;
input models $20. ;
datalines;
  Linear
  LogLinear
  Emax
  SigEmax
  Exponential
  Logistic
  Quadratic
  BetaModel
run;

data dose;
input dose n;
datalines;
  00 20
  10 20
  20 20
  ...
run;

```

After that, initial values of the model parameters in nonlinear candidate models need to be defined. Note that the Linear and Log-Linear models do not require any initial guesses. The initial parameter values are obtained by calling the model-specific macros, as explained below. Note that some dose-response models, e.g., the logistic and SigEmax models, have the same input pairs due to similar shape.

- `%Emax` macro with two parameters (`d` and `p`). This macro calculates the guess estimates of the Emax model parameters, i.e.,  $ED_{50}$  and  $E_{\max}$ , for a range of paired values of `d` (dose) and `p` (prior expected probability of the maximum effect). To guess  $ED_{50}$ , a dose  $d$  is considered that maximizes the treatment effect with the probability  $p$ . As an illustration, assume that a dose of 100 mg has a 75% probability of achieving the highest potency ( $E_{\max}$ ). The parameter  $ED_{50}$  is estimated from  $p = d/(ED_{50} + d)$ . The resulting guess estimate of  $ED_{50}$  is 33.3 mg.
- `%Exponential` macro with two parameters (`d` and `p`). The initial guess assumes that for a dose  $d$ , the expected increase in the treatment effect has the probability  $p$ . The equation  $1 - e^{d/\delta} = p(1 - e^{D/\delta})$ , where  $D$  is the maximum dose, is used to find  $\delta$  by applying a bisection method in this SAS macro.
- `%Logistic` macro with four parameters (`d1, p1, d2` and `p2`). Enter the probabilities

$p_1$  and  $p_2$  such that the two doses  $d_1$  and  $d_2$  estimate  $E_{max}$ . The resulting estimate of  $ED_{50}$  is given by

$$\frac{d_1 \text{logit}(p_2) - d_2 \text{logit}(p_1)}{\text{logit}(p_2) - \text{logit}(p_1)}.$$

- **%SigEmax** macro with four parameters (`d1`, `p1`, `d2`, and `p2`). As in the Emax model, to guess  $ED_{50}$ , select a dose  $d_1$  that maximizes the treatment effect  $E_{max}$  with the probability  $p_1$ . To estimate the Hill factor ( $h$ ), propose a second pair of dose-probability values ( $d_2$  and  $p_2$ ) that maximizes the treatment effect  $E_{max}$ . The Hill factor is estimated as

$$\frac{\log(p_1(1-p_2)/(p_2(1-p_1)))}{\log(d_1/d_2)}$$

and  $ED_{50}$  is estimated as

$$d_1 \left( \frac{1-p_1}{p_1} \right)^{1/h}.$$

- **%Quadratic** macro with two parameters (`d` and `p`). The initial guess of the dose  $d$  assumes a treatment effect of  $ED_{50}$  with the probability  $p$  and is solved using a quadratic equation.
- **%BetaModel** macro with four parameters (`d`, `p`, `dMax`, and `Scal`). The initial guess of the dose  $d$  has the probability of  $p$  to be  $ED_{50}$ . The kernel of the beta model function consists of the kernel of the density function of a beta distribution on the interval  $[0, S]$ . The  $S$  parameter (`Scal`) is set at a default value of 1.2, which implies 20% more than the maximum dose. The initial value of `dMax`, i.e., the dose at which the maximum effect occurs, should be defined to ensure that the `Scal` parameter is greater than `dMax`.

A macro from this list needs to be executed only if the corresponding model is included in the candidate set, i.e., it is specified in the `models` data set.

To finalize the set up, the **%StartUp** macro needs to be called. This macro creates and populates macro variables that will be used in the dose-finding algorithm.

Optimal contrasts are identified in Step 2 using the following macros:

- **%RegularOptimalContrasts** computes the optimal dose-response contrasts based on simple candidate models without covariates. This macro will be illustrated using Case study 1.
- **%GeneralizedOptimalContrasts** computes the optimal dose-response contrasts based on candidate models with covariate adjustment. This macro will be applied to more complex settings based on Case study 2.

Both macros can be used with general unbalanced designs and dosing schemes with unequally spaced dose levels.

The adjusted critical value in Step 3 is computed by invoking the **%CriticalValue** macro that supports parametric multiplicity adjustments for dose-response contrasts. The remaining two steps are performed using appropriate SAS procedures. See Sections 4.4 and 4.5 for details.

### 4.3.5 Summary

Several testing strategies commonly used in dose-finding Phase II trials were described in this section. This includes simple pairwise tests that are carried out in all confirmatory Phase III trials with multiple dose-control comparisons.

The pairwise tests tend to be underpowered in Phase II trials, and it is recommended to use contrast-based tests instead. Contrast-based tests are aimed at evaluating the overall dose-related trend. A contrast-based test performs best if its coefficients mimic the shape of the dose-response function. Since the underlying dose-response function is unknown, it is most sensible to consider multi-contrast tests that rely on a family of predefined contrasts that correspond to a set of plausible dose-response functions. Multi-contrast tests serve as robust tools for dose-response analysis within the hypothesis testing framework.

The MCP-Mod procedure was introduced as an extension of multi-contrast tests. This procedure simultaneously addresses the goals of hypothesis testing and estimation in dose-finding trials. As a result, the general approach implemented within this procedure supports all three components of dose-response analysis in Phase II trials. This includes the assessment of the dose-response signal based on powerful contrast-based tests; identification of the best model to estimate the dose-response relationship; and, finally, determination of target doses based on the estimated dose-response function. The MCP-Mod dose-finding algorithm will be illustrated using Case studies 1 and 2 in Sections 4.4 and 4.5. It will be shown in these sections how dose-response analysis can be efficiently performed in clinical trials with a variety of primary endpoints and dose-response modeling strategies, including strategies that account for missing data.

## 4.4 Dose finding in Case study 1

---

The discussion of dose-finding approaches in this case study will rely on the MCP-Mod procedure introduced in Section 4.3.4. Section 4.4.1 will focus on a simple setting with a predefined set of candidate models without covariates. A more challenging case that considers models with several covariates will be presented in Section 4.4.2.

### 4.4.1 MCP-Mod procedure based on simple ANOVA models

The presentation of the MCP-Mod dose-finding algorithm in the schizophrenia trial from Case study 1 will follow the five steps defined in Section 4.3.4.

#### Step 1: Candidate models

As explained in Section 4.3.4, an exploration of the dose-response relationships within the MCP-Mod framework begins with a specification of candidate dose-response models. As an illustration, four models (Linear, LogLinear, Emax, and Exponential) will be included in the candidate set in this clinical trial example. A joint assessment of the four models is expected to provide a full interpretation of the dose response's shape in the schizophrenia trial.

Program 4.3 specifies the candidate set of models (they are included in the `models` data set) and provides information about the dose levels that are investigated in the trial (0 mg, 40 mg, 80 mg, and 120 mg) and number of patients in each arm (60 patients per arm). This information is specified in the `dose` data set. The `%StartUp` macro is invoked to set up macro variables that will be used later in the dose-finding algorithm.

The next step is to determine the initial values of model parameters for the dose-response models in the candidate set. These initial values will be used in the calculation of optimal contrasts in Step 2. As explained before, no start values are

needed if the model is based on a linear or log-linear fit, but initial values are required for nonlinear models, i.e., the Emax and Exponential models in this example. The initial values of model parameters in the two models are computed at the bottom of Program 4.3 using the rules defined at the end of Section 4.3.4. Beginning with the Exponential model, the  $\delta$  parameter controls the convexity of the dose-response model, and the initial value of  $\delta$  is defined by the `delta` macro variable, which is set to 1 in this program. The start value of the maximum treatment effect ( $E_{\text{max}}$ ) is defined in the Emax model using the `MaxEff` macro variable. This variable is set to 1 or, in other words, 100%. An offset parameter ( $c$ ) needs to be specified in the LogLinear model to avoid issues with the zero dose. The offset parameter is set to 0.1 in this case, using the `offset` macro variable. The `%Emax` and `%Exponential` macros are called to obtain the initial guesses in the corresponding dose-response models. In both cases, the initial values are computed under the assumption that the 100-mg dose provides a 75% improvement over placebo with respect to the mean reduction from baseline in the PANSS total score. As a result, both macros are invoked with `d=100` and `p=0.75`.

### **PROGRAM 4.3 Candidate models in Case study 1**

```

data models;
input models $20. ;
datalines;
    Linear
    LogLinear
    Emax
    Exponential
run;

data dose;
input dose n;
datalines;
    0 60
    40 60
    80 60
    120 60
run;

%let delta=1;
%let offset=0.1;
%let MaxEff=1;

%StartUp;

%Emax(d=100, p=0.75);
%Exponential(d=100, p=0.75);

```

The actual values of the initial model parameters are not important, but, for the reader's general information, the resulting guesstimate of the  $ED_{50}$  parameter in the Emax model is 33.3, and the  $\delta$  parameter is assumed to equal 110.4 in the Exponential model.

### **Step 2: Optimal contrasts**

With the initial estimates of model parameters stored in appropriate macro variables, the next step is to derive the optimal dose-response contrasts. Program 4.4 executes

the `%RegularOptimalContrasts` macro to calculate the sets of model-specific contrast coefficients as well as the associated correlation matrix. This particular macro is invoked since the candidate models do not include co-variates. The macro relies on the macro variables created in Step 1, and no arguments need to be specified.

#### PROGRAM 4.4 Optimal contrasts in Case study 1

```
%RegularOptimalContrasts;
```

It will be shown in Section 4.4.2 that the `%GeneralizedOptimalContrasts` macro needs to be called to perform dose-finding in the presence of baseline covariates.

The optimal contrast coefficients generated by the `%RegularOptimalContrasts` macro are listed in Table 4.5. These coefficients are saved in the `optcont` data set to be used at later stages of the dose-finding algorithm.

**TABLE 4.5 Output from Program 4.4: Model-specific optimal contrast coefficients**

Dose	Placebo	40 mg	80 mg	120 mg
Linear	-0.671	-0.224	0.224	0.671
LogLinear	-0.807	-0.005	0.307	0.505
Emax	-0.831	0.061	0.323	0.448
Exponential	-0.587	-0.291	0.134	0.744

Program 4.4 also computes the correlation matrix based on these contrasts. The matrix is shown in Table 4.6. It was noted in Section 4.3.4 that this correlation matrix only depends on the design parameters such as the sample sizes in the individual trial arms. It follows from this table that some of the contrasts are strongly correlated with each other. For example, the correlation between the contrasts derived from the Linear and Emax models exceeds 99%. Similarly, there is a very strong correlation between the contrasts based on the Linear and Exponential models. This indicates that the shapes of the corresponding dose-response models are very close to each other, and, thus, they may potentially be redundant. The correlation matrix generated by Program 4.4 is saved in the `corrmat` data set and will be used in Step 3 to find a multiplicity-adjusted critical value for the test statistics associated with the four optimal dose-response contrasts in this clinical trial example.

**TABLE 4.6 Output from Program 4.4: Correlations among optimal contrasts**

	Linear	LogLinear	Emax	Exponential
Linear	1.000	0.950	0.917	0.988
LogLinear	0.950	1.000	0.996	0.891
Emax	0.917	0.996	1.000	0.847
Exponential	0.988	0.891	0.847	1.000

It is worth pointing out that the numerical values shown in Tables 4.5 and 4.6 are rounded off to three decimal places. This rounding was necessary to facilitate presentation. However, the original values are used at the subsequent stages of the dose-finding algorithm. Using rounded values instead of the original values can have a significant impact on the conclusions, especially if the optimal contrasts are strongly correlated.

#### Step 3: Dose-response tests

To assess the significance of the dose-response relationship based on each optimal contrast, Program 4.4 uses PROC SQL to save each set of contrast coefficients to a

macro variable. For example, the `linear_cont` macro variable represents a string of optimal contrast coefficients derived from the Linear model. The resulting macro variables are then passed to PROC GLM to carry out the simple *t* test for each dose-response contrast.

#### **PROGRAM 4.5 Dose-response tests based on optimal contrasts in Case study 1**

```
proc sql;
    select Linear into :linear_cont separated by " "
        from optcont;
    select LogLinear into :loglinear_cont separated by " "
        from optcont;
    select Emax into :emax_cont separated by " "
        from optcont;
    select Exponential into :exponential_cont separated by " "
        from optcont;

proc glm data=cs1;
    class dose;
    model improvement=dose;
    Estimate 'Linear'      dose &linear_cont;
    Estimate 'LogLinear'   dose &loglinear_cont;
    Estimate 'Emax'        dose &emax_cont;
    Estimate 'Exponential' dose &exponential_cont;
run;
```

The output of Program 4.5 is displayed in Table 4.7. It is easy to verify that the test statistics listed in this table are all significant at a one-sided  $\alpha = 0.025$ , which was predefined in this case study. However, it is premature to interpret this as evidence of a truly significant dose-response trend since we have not accounted for the multiplicity induced by the simultaneous investigation of four models.

**TABLE 4.7 Output from Program 4.5**

Model	Contrast estimate	Standard error	Test statistic
Linear	4.118	1.7875	2.30
LogLinear	4.044	1.7875	2.26
Emax	3.938	1.7875	2.20
Exponential	3.872	1.7875	2.17

To select the clinically relevant dose-response models, i.e., the models associated with significant contrasts, the multiplicity-adjusted critical value that corresponds to a one-sided  $\alpha = 0.025$  needs to be calculated. The multiplicity-adjusted critical value is computed from a central multivariate *t* distribution by calling the `%CriticalValue` macro with the following parameters:

- `corrmat` is the correlation matrix.
- `nu` is the number of degrees of freedom.
- `alpha` is the one-sided Type I error rate.

It is important to note that this macro takes advantage of the powerful MVT function written by Frank Bretz and Alan Genz. This function evaluates multivariate *t* probabilities using the method developed by Genz and Bretz (2002).

In this clinical trial example, the correlation matrix of the multivariate *t* distribution is shown in Table 4.6 and can be passed to the macro using the `corrmat` data set. The number of degrees of freedom is given by  $\nu = N - m = 236$  since

$N = 240$  is the total number of patients and  $m = 4$  is the number of trial arms. Lastly,  $\alpha=0.025$ . Program 4.6 calls the %CriticalValue macro to find the multiplicity-adjusted critical value for the four optimal contrasts.

#### PROGRAM 4.6 Multiplicity-adjusted critical value in Case study 1

```
%CriticalValue(corr=corrmat, nu=236, alpha=0.025);
```

The multiplicity-adjusted critical value computed by Program 4.6 is equal to  $q = 2.155$ . An iterative algorithm is employed to find this critical value, and the process might take a few minutes.

We can see from Table 4.7 that all four test statistics are greater than this adjusted critical value, and, thus, all four models result in a significant dose-response relationship after the multiplicity adjustment. The final set of clinically relevant models includes the Linear, LogLinear, Emax, and Exponential models.

#### Step 4: Model selection

Focusing on the final set of dose-response models, there are several approaches to select the best model as outlined in Step 4 of the MCP-Mod algorithm. In this example, we will consider model selection rules based on the maximum contrast test statistic and commonly used information criteria such as the AIC and BIC.

Note that several different procedures are available in SAS/STAT to fit linear and nonlinear models and compute a variety of information criteria. The procedures differ in the underlying assumptions and fit criteria. It is advisable to consider the procedures that fit all of the predefined dose-response models, both linear and nonlinear models, to make sure that the selected information criteria are evaluated in a consistent way. Most of the time, it is convenient to use PROC NLMIXED because it supports a very broad class of nonlinear models. Program 4.7 runs this procedure to fit the four dose-response models. For the nonlinear model parameter estimation, initial values are necessary, and convergence from these starting values might lead to local minima. An educated initial values coupled with multi-point testing tends to yield consistent results. As before, initial values are not necessary for the Linear and LogLinear models. For the LogLinear model, an offset is defined as 10% of the maximum dose in the trial, i.e., 12. For the Emax and Exponential models, the initial value of  $e_0$  is guessed from the intercept of the LogLinear model, i.e., 5. Further, the  $E_{max}$  parameter in the Emax model is the maximum expected treatment effect, and it is reasonable to set it to 15. The  $e_1$  parameter for the Exponential model can be tricky to guess. For this reason, no initial value is specified in Program 4.7. The  $\delta$  parameter is set to two times the maximum dose, i.e., 240.

#### PROGRAM 4.7 Parameter estimation in clinically relevant models in Case study 1

```
* Linear model;
proc nlmixed data=cs1;
  eta=e0+delta*dose;
  model improvement~normal(eta, s2);
  ods output ParameterEstimates=LinParameterEstimates
    FitStatistics=AIC1(rename=(Value=AIC)
      where=(Descr='AIC (smaller is better)' ));
run;

* LogLinear model;
proc nlmixed data=cs1;
```

```

eta=e0+delta*log(dose+12);
model improvement~normal(eta, s2);
ods output ParameterEstimates=LogLinParameterEstimates
FitStatistics=AIC2(rename=(Value=AIC)
where=(Descr='AIC (smaller is better)'));

run;

* Emax model;
proc nlmixed data=cs1;
parms e0=5 emax=15;
y=e0 + (emax*dose) / (ED50+dose);
model improvement~normal(y,exp(s2));
ods output ParameterEstimates=EmaxParameterEstimates
FitStatistics=AIC3(rename=(Value=AIC)
where=(Descr='AIC (smaller is better)'));

run;

* Exponential model;
proc nlmixed data=cs1;
parms e0=5 delta=240;
y= e0 + e1 * (exp(dose/delta)-1);
model improvement~normal(y,exp(s2));
ods output ParameterEstimates=ExpParameterEstimates
FitStatistics=AIC4(rename=(Value=AIC)
where=(Descr='AIC (smaller is better)'));

run;

data aic;
set aic1 aic2 aic3 aic4;
run;

```

The four dose-response models fitted in Program 4.7 are plotted in Figure 4.4. We can see from this figure that the general shapes of the dose-response functions based on the Linear, Emax, and Exponential models are quite similar to each other. This is due to the fact that the Emax and Exponential functions are almost linear over the selected dose range. The LogLinear has a concave shape and clearly underestimates the mean effect in the placebo arm.

Program 4.7 also generates the `aic` data set that includes the AIC values that correspond to the four dose-response models in the final set. The same code can be used to compute the BIC values. To accomplish this, the condition

```
where=(Descr='AIC (smaller is better)')
```

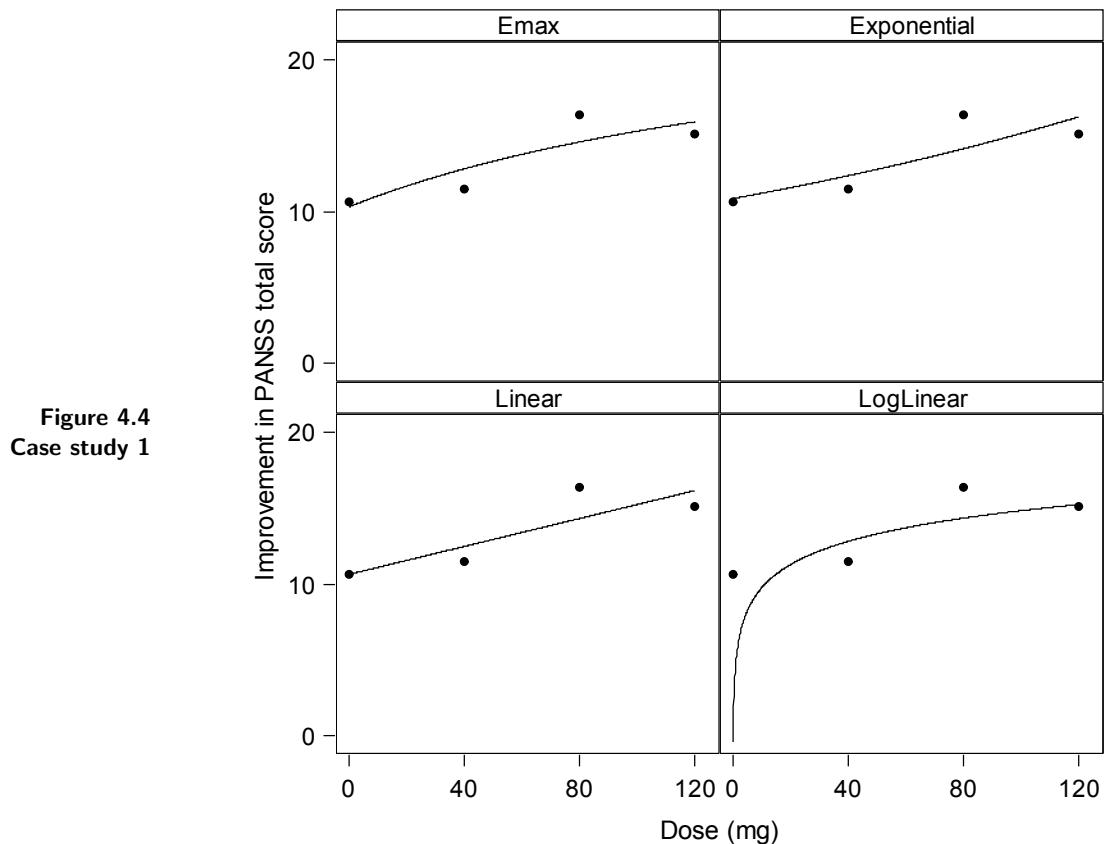
needs to be replaced with

```
where=(Descr='BIC (smaller is better)')
```

The resulting AIC and BIC values are displayed in Table 4.8. This table also displays the parameters of the dose-response models plotted in Figure 4.4.

**TABLE 4.8 Dose-response models based on optimal contrasts**

Model	Parameter estimates	Test statistic	AIC	BIC
Linear	$E_0 = 10.63, \delta = 0.046$	2.30	1946.5	1956.8
LogLinear	$E_0 = 4.64, \delta = 2.212$	2.26	1946.6	1957.0
Emax	$E_0 = 10.31, E_{\max} = 14.75, ED_{50} = 196.25$	2.20	1948.2	1962.1
Exponential	$E_0 = 10.85, E_1 = 8.3, \delta = 239.89$	2.17	1948.7	1962.6



**Figure 4.4**  
Case study 1

*Dose-response models based on optimal contrasts.*

As pointed out at the beginning of this subsection, the best dose-response model can be identified using selection rules based on the contrast test statistics or information criteria.

Beginning with the model selection rule that relies on the maximum contrast test statistic, it is shown in Table 4.8 that the test statistic of the Linear contrast ( $T_1 = 2.30$ ) is greater than the other statistics. This means that the Linear model should be chosen as the best dose-response model. An important limitation of this model selection rule is that the number of model parameters is not taken into account. For example, if the test statistic associated with the Emax contrast was the most significant statistic, we could argue that this might be the result of overfitting since the number of parameters in the Emax model is greater than the number of parameters in the Linear model. The simple approach that relies on the maximum contrast test statistic does not incorporate a penalty for the number of model parameters.

Model selection rules based on information criteria help address this limitation of the simple rule discussed above. The AIC and BIC are designed to take a model's complexity (number of parameters) into account and thus tend to serve as more reliable tools for model selection. With either information criterion, the best model corresponds to the lowest value of the criterion. Using the AIC as an example, the Linear model with  $AIC_1 = 1946.5$  is chosen as the best model. The same conclusion is reached if the BIC is considered. However, in general, the rules based on the two criteria might not be consistent and it might be difficult to justify the selection of a single model that provides the best fit to the data.

## Step 5: Target dose selection

Based on the final set of dose-response models defined in Table 4.8, we are ready to discuss the important topic of dose selection based on the data in the schizophrenia trial. Recall from Step 5 of the MCP-Mod algorithm that the Minimum Effective Dose (MED) is defined as the lowest dose that still provides a clinically meaningful and statistically significant improvement over placebo. To estimate this dose in the schizophrenia trial, we need to specify the threshold, denoted by  $\Delta$ , that determines the clinically relevant effect. We will assume that the treatment difference of 5 points on the PANSS Total Score scale compared to the placebo arm is minimally clinically meaningful and thus  $\Delta = 5$ .

Program 4.8 computes the minimum effective doses for each of the four clinically relevant dose-response models identified in Step 3. As an illustration, consider the MED estimation for the LogLinear model by inverting this nonlinear function. Since the difference in the treatment effect between the MED dose and placebo is 5, we have

$$[E_0 + \delta \log(\text{MED} + c)] - [E_0 + \delta \log(0 + c)] = 5.$$

It follows from this equation that

$$\text{MED} = c(e^{5/\delta} - 1).$$

Similar approaches are applied to estimate the MED based on the other nonlinear models in this case study. The parameter values used in this program are based on the estimates listed in Table 4.8.

### PROGRAM 4.8    Minimum effective doses computed from clinically relevant models in Case study 1

```

data LinParameterEstimates_;
  merge LinParameterEstimates(where=(Parameter='delta')) AIC1;
    MED_Lin=5/Estimate;
    rename AIC=AIC_Lin;
run;

proc transpose data=LogLinParameterEstimates out=LogLinParameterEstimates_;
  var Estimate;
  id Parameter;
run;

data LogLinParameterEstimates_;
  merge LogLinParameterEstimates_ AIC2;
    MED_LogLin=12*exp(5/delta)-12;
    rename AIC=AIC_LogLin;
run;

proc transpose data=EmaxParameterEstimates out=EmaxParameterEstimates_;
  var Estimate;
  id Parameter;
run;

data EmaxParameterEstimates_;
  merge EmaxParameterEstimates_ AIC3;
    MED_Emax=5*ED50/(emax-5);
    rename AIC=AIC_Emax;
run;

```

```

proc transpose data=ExpParameterEstimates out=ExpParameterEstimates_;
  var Estimate;
  id Parameter;
run;

data ExpParameterEstimates_;
  merge ExpParameterEstimates_ AIC4;
    MED_Exp=delta*log(1+5/e1);
    rename AIC=AIC_Exp;
run;

data med;
  merge LinParameterEstimates_ LogLinParameterEstimates_
    EmaxParameterEstimates_ ExpParameterEstimates_;
  keep AIC_Lin MED_Lin
    AIC_LogLin MED_LogLin
    AIC_Emax MED_Emax
    AIC_Exp MED_Exp;
run;

```

The model-specific minimum effective doses computed by Program 4.8 are listed in Table 4.9. If the Linear model is chosen as the best dose-response model using the arguments presented in Step 4, the final MED in this trial is estimated as 108.6 mg.

**TABLE 4.9 Output from Program 4.8**

Model	MED (mg)
Linear	108.6
LogLinear	103.1
Emax	100.7
Exponential	113.1

It is important to note that the target dose selection might not be consistent across the clinically relevant models. For example, as shown in Table 4.9, the estimated MED ranges from 100.7 mg to 113.1 mg. To develop a unified dose selection strategy, it is recommended to consider the model averaging approach described in Step 5 of the MCP-Mod algorithm. This approach results in a single estimate of the target dose by averaging over an appropriate information criterion. To accomplish this, the prior probabilities, denoted by  $p_1$  through  $p_4$ , need to be assigned to the four dose-response models. These probabilities help quantify the strength of prior information to support each model. In this particular case, the non-informative prior probabilities given by  $p_1 = \dots = p_4 = 0.25$  will be used.

For this example, model averaging will be illustrated using the AIC. Note, first, that a direct application of the formula for computing model-specific weights introduced in Step 5 might not be optimal from a computational perspective. Based on the formula, the weight for the Linear model is

$$w_1 = \frac{p_1 \exp(-\text{IC}_1/2)}{\sum_{j=1}^4 p_j \exp(-\text{IC}_j/2)}.$$

For the numerator, a value of  $\exp(-1946.5/2)$  might be mathematically non-zero, but it will be approximated by zero in SAS or most other software packages. A simple way to address this problem is to make a minor adjustment to the formula and define the weight of the first model as follows:

$$w_1 = \frac{p_1}{\sum_{j=1}^4 p_j \exp(\text{IC}_1/2 - \text{IC}_j/2)}.$$

Using this formula, the weight of the Linear model based on the AIC is easily computed as  $w_1 = 0.379$ . The remaining model weights are calculated in a similar way. Program 4.9 computes the model-specific weights based on the AIC using the `med` data set created in Program 4.8. This program is easily modified to perform calculations using any other information criterion, including the BIC.

#### **PROGRAM 4.9    Model-specific weights based on AIC in Case study 1**

```

data weighted_med;
set med;
p=1/4;
array x[*] AIC_Lin AIC_LogLin AIC_Emax AIC_Exp;
array A[*] A1-A4;
array B[*] B1-B4;
array C[*] C1-C4;
array D[*] D1-D4;
do i=1 to dim(x);
A{i}=x{i}-x{1};
B{i}=x{i}-x{2};
C{i}=x{i}-x{3};
D{i}=x{i}-x{4};
end;
Wt1=p/(p*(exp(-A1/2)+exp(-A2/2)+exp(-A3/2)+exp(-A4/2)));
Wt2=p/(p*(exp(-B1/2)+exp(-B2/2)+exp(-B3/2)+exp(-B4/2)));
Wt3=p/(p*(exp(-C1/2)+exp(-C2/2)+exp(-C3/2)+exp(-C4/2)));
Wt4=p/(p*(exp(-D1/2)+exp(-D2/2)+exp(-D3/2)+exp(-D4/2)));

weighted_med=Wt1*MED_Lin+Wt2*MED_LogLin+Wt3*MED_Emax+Wt4*MED_Exp;
keep Wt1-Wt4 weighted_med;
run;

```

The model-specific weights based on two popular information criteria (AIC and BIC) are listed in Table 4.10. Based on these weights, we observe that the Linear model is given prominence while the Exponential model is weighed the least. This is consistent with our understanding that a higher weight should be assigned to a model with a lower AIC or BIC value.

**TABLE 4.10    Model-specific weights based on AIC and BIC**

Model	Model's weight	
	AIC	BIC
Linear	0.379	0.491
LogLinear	0.344	0.446
Emax	0.154	0.035
Exponential	0.123	0.028

Program 4.9 also computes the minimum effective dose based on the model averaging method. This dose is computed using the model-specific weights shown in Table 4.10 as follows:

$$\widehat{\text{MED}} = \sum_{j=1}^4 w_j \widehat{\text{MED}}_j.$$

This results in the MED of 106 mg based on the AIC. Similarly, using the model weights derived from BIC, the MED is also 106 mg. The optimal doses computed from the weighted model are marginally different from the optimal dose based on the single best model, i.e., the Linear model, which is equal to 108.6 mg (see Table 4.9). The information on MEDs based on the individual dose-response models

and weighted models will need to be used along with clinical judgment and other relevant information to select the MED for the experimental treatment in Case study 1.

#### 4.4.2 MCP-Mod procedure based on ANCOVA models

In Section 4.4.1, we defined a dose-finding algorithm based on the MCP-Mod procedure using a simple setting with four one-way ANOVA models. The models did not include any baseline covariates. This section presents an extended MCP-Mod dose-finding algorithm for a more realistic setting where several potential prognostic covariates (i.e., covariates that are likely to affect the outcome variable) are predefined. In a general clinical trial setting, it is important to consider covariates that can influence the outcome variable of interest. If these prognostic covariates are ignored, the sponsor is likely to arrive at incorrect conclusions. Commonly used covariates include the baseline value of the outcome variable, gender, disease severity, etc. In this case study, we will consider a family of dose-response models that account for the potential prognostic effect of the two baseline covariates:

- Gender
- Baseline PANSS total score

The MCP-Mod procedure based on the simple ANOVA model used in Section 4.4.1 is easily extended to a variety of more complex settings. See, for example, Pinheiro et al. (2013). It will be shown below how the dose-finding algorithm is applied when candidate dose-response models include covariates.

#### Step 1: Candidate models

Dose-response models will be constructed for the improvement in the PANSS total score with an adjustment for the selected covariates. In this case, the generalized linear modeling approach is applied, and the primary endpoint is assumed to depend on the dose through the parameter  $\mu = \mu(d)$ , i.e., the distribution of the outcome variable  $y$  is defined as follows

$$y \sim F(x, \nu, \mu(d)),$$

where  $x$  denotes the vector of covariates and  $\nu$  is a vector or a nuisance parameter. We follow the steps similar to those used in the MCP-Mod dose-finding algorithm defined in Section 4.4.1 and focus on the parameter  $\mu(d)$ , its estimate  $\hat{\mu}(d)$ , and the covariance matrix of  $\hat{\mu}(d)$  denoted by  $S$ . This approach can be applied to set up a generalized MCP-Mod procedure based on any of the dose-response models listed in Table 4.4. To illustrate dose finding based on ANCOVA models, the four candidate models used in Section 4.4.1 (i.e., Linear, LogLinear, Emax and Exponential models) will be considered in this setting.

#### Step 2: Optimal contrasts

The computation of optimal covariate-adjusted contrasts is supported by the `%GeneralizedOptimalContrasts` macro. This macro is called in Program 4.10.

**PROGRAM 4.10 Optimal covariate-adjusted contrasts in Case study 1**

```

proc genmod data=cs1 order=data;
  class dose gender subjid;
  model improvement=dose gender baseline /link=identity dist=normal noint;
  repeated subject=subjid/type=ind covb;
  ods output GEENCov=S(drop=RowName)
    GEEEmpPEst=muHat(where=(Parm=(‘dose’)));
run;

* Select Parameters 2 through 5 only (dose effects);
data S(where =(x<5));
  set S;
  keep x prm2-prm5;
  x=_n_;
run;

data S;
  set S;
  drop x;
run;

%GeneralizedOptimalContrasts;

```

Program 4.10 finds an estimate of the parameter  $\mu(d)$ , which is saved in the `muHat` data set, and its covariance matrix  $S$ . To extract this covariance matrix, note that PROC GENMOD computes the covariance matrix for the six parameters in the simple linear model with the two baseline covariates. These parameters are labeled Prm2, Prm3, Prm4, Prm5, Prm6, and Prm8 and correspond to the four trial arms and two covariates. This  $6 \times 6$  covariance matrix is saved in the `S` data set. Since our focus is the estimation of the dose-related effects, the covariance matrix  $S$  is obtained by removing the last two parameters. This resulting covariance matrix is given by

$$S = \begin{bmatrix} 62.96 & 59.82 & 57.15 & 57.46 \\ 59.82 & 62.85 & 57.09 & 57.41 \\ 57.15 & 57.09 & 57.61 & 54.84 \\ 57.46 & 57.41 & 54.84 & 58.21 \end{bmatrix}.$$

The optimal contrasts adjusted for the covariate effects are presented in Table 4.11. The contrast coefficients can be found in the `optcont` data set generated by the `%GeneralizedOptimalContrasts` macro.

**TABLE 4.11 Output from Program 4.10: Model-specific covariate-adjusted optimal contrast coefficients**

Dose	Placebo	40 mg	80 mg	120 mg
Linear	-0.672	-0.216	0.213	0.675
LogLinear	-0.810	0.007	0.300	0.504
Emax	-0.835	-0.073	0.316	0.445
Exponential	-0.585	-0.285	0.121	0.749

It is instructive to compare these optimal contrasts with the optimal contrasts based on simple dose-response models without covariate adjustment. Comparing the contrast coefficients listed in Tables 4.5 and 4.11, we see that the two covariates appear to have little impact on the contrast coefficients.

Using the covariance matrix  $S$ , the `%GeneralizedOptimalContrasts` macro in Program 4.10 computed the correlation matrix for the covariate-adjusted optimal contrasts. This matrix is saved in the `corrmat` data set and is shown in Table 4.12.

**TABLE 4.12 Output from Program 4.10: Correlations among covariate-adjusted optimal contrasts**

	Linear	LogLinear	Emax	Exponential
Linear	1.000	0.947	0.913	0.987
LogLinear	0.947	1.000	0.996	0.886
Emax	0.913	0.996	1.000	0.840
Exponential	0.987	0.886	0.840	1.000

### Step 3: Dose-response tests

To identify the contrasts that are associated with a significant dose-response relationship, the  $t$  statistics for the four optimal contrasts were computed using PROC GLM. See Program 4.11. As in Program 4.5, PROC SQL was used to create macro variables with the contrast coefficients.

**PROGRAM 4.11 Dose-response tests based on optimal covariate-adjusted contrasts in Case study 1**

```

proc sql;
    select Linear into :linear_cont separated by " "
        from optcont;
    select LogLinear into :loglinear_cont separated by " "
        from optcont;
    select Emax into :emax_cont separated by " "
        from optcont;
    select Exponential into :exponential_cont separated by " "
        from optcont;

proc glm data=cs1;
    class dose gender subjid;
    model improvement=dose gender baseline;
    Estimate 'Linear'      dose &linear_cont;
    Estimate 'LogLinear'   dose &loglinear_cont;
    Estimate 'Emax'        dose &emax_cont;
    Estimate 'Exponential' dose &exponential_cont;
run;

```

It needs to be pointed out that the effect of the baseline PANSS total score on the improvement to the end of the 4-week treatment period was highly statistically significant, but the effect of gender was not significant across the models. As anticipated, patients who had higher PANSS total scores to begin with were likely to experience a greater improvement in their PANSS total score over the treatment period, and patients with lower baseline scores did not, in general, improve that much. The output of Program 4.11 is displayed in Table 4.13.

**TABLE 4.13 Output from Program 4.11**

Model	Contrast estimate	Standard error	Test statistic
Linear	5.057	1.7767	2.85
LogLinear	4.908	1.7740	2.77
Emax	4.758	1.7724	2.68
Exponential	4.768	1.7751	2.69

We see that the contrast test statistics listed in Table 4.13 are larger in magnitude than the test statistics that were computed from the four unadjusted models (see Table 4.7). For example, the most significant statistic in the unadjusted

case (2.30) is much lower than the least significant statistic shown in Table 4.13 (2.68). This increase is directly due to the fact that the candidate models accounted for the effect of an important prognostic variable (PANSS total score at baseline).

Program 4.12 computes the multiplicity-adjusted critical value to select clinically meaningful dose-response models from the candidate set. This critical value is found by calling the same macro (%CriticalValue macro) as in the case of unadjusted dose-response models. The correlation matrix derived in Program 4.10 is passed to the %CriticalValue macro along with two other parameters. The number of degrees of freedom (nu) is defined as in Section 4.4.1 since it depends only on the total sample size in the trial and number of arms. Similarly, the one-sided Type error rate (alpha) is again set to 0.025.

#### **PROGRAM 4.12 Multiplicity-adjusted critical value in Case study 1**

```
%CriticalValue(corr=corrmat, nu=236, alpha=0.025);
```

The adjusted critical value is equal to 2.159. A quick review of the model-specific test statistics listed in Table 4.13 reveals that all of them are significant after the multiplicity adjustment. As a consequence, all four models are included in the final set of clinically relevant dose-response models.

#### **Step 4: Model selection**

After the clinically relevant dose-response models have been identified, the model parameters need to be estimated. As discussed earlier, it is advisable to consider the SAS procedure that fits all of the candidate models (both linear and nonlinear models) to make sure that comparable values of appropriate information criteria (e.g., the AIC or BIC) are used. We estimated the parameters of the four selected models with an adjustment for the baseline covariates using PROC NLMIXED in Program 4.13.

#### **PROGRAM 4.13 Parameter estimation in clinically relevant covariate-adjusted models in Case study 1**

```
* Linear model;
proc nlmixed data=cs1;
  eta=e0+delta*dose+beta1*baseline+u;
  model improvement~normal(eta, s2);
  random u ~normal(0, s2u) subject=subjid;
  ods output ParameterEstimates=LinParameterEstimates
    FitStatistics=AIC1(rename=(Value=AIC)
      where=(Descr='AIC (smaller is better)' ));
run;

* LogLinear model;
proc nlmixed data=cs1;
  eta=e0+delta*log(dose+12)+beta1*baseline+u;
  model improvement~normal(eta, s2);
  random u ~normal(0, s2u) subject=subjid;
  ods output ParameterEstimates=LogLinParameterEstimates
    FitStatistics=AIC2(rename=(Value=AIC)
      where=(Descr='AIC (smaller is better)' ));
run;
```

```

* Emax model;
proc nlmixed data=cs1;
  parms e0=5 emax=15;
  y=e0 + (emax*dose) / (ED50+dose)+beta1*baseline+u;
  model improvement~normal(y,exp(s2));
  random u ~normal(0, s2u) subject=subjid;
  ods output ParameterEstimates=EmaxParameterEstimates
    FitStatistics=AIC3(rename=(Value=AIC)
    where=(Descr='AIC (smaller is better)'));

run;

* Exponential model;
proc nlmixed data=cs1;
  parms e0=-15 delta=240 e1=10;
  y= e0 + e1 * (exp(dose/delta)-1)+beta1*baseline+u;
  model improvement~normal(y,exp(s2));
  random u ~normal(0, s2u) subject=subjid;
  ods output ParameterEstimates=ExpParameterEstimates
    FitStatistics=AIC4(rename=(Value=AIC)
    where=(Descr='AIC (smaller is better)'));

run;

```

The estimated parameters of the four models and associated AIC values are shown in Table 4.14. It follows from the table that the information criterion is minimized if the Linear model is selected to describe the dose-response relationship in the schizophrenia clinical trial. The model selection rule based on the maximum test statistic also identifies the Linear model as the best model in this setting.

**TABLE 4.14** Dose-response models based on covariate-adjusted optimal contrasts

Model	Parameter estimates	Test statistic	AIC
Linear	$E_0 = -14.29, \delta = 0.057$	2.85	1940.0
LogLinear	$E_0 = -21.22, \delta = 2.688$	2.77	1940.4
Emax	$E_0 = -15.11, E_{\max} = 15.61, ED50 = 138$	2.68	1941.8
Exponential	$E_0 = -14.69, E_1 = 10.41, \delta = 239.99$	2.69	1942.3

### Step 5: Target dose selection

The selection of the minimum effective dose will be performed using the same approach as in Section 4.4.1, i.e., the clinically relevant improvement over placebo will be set to  $\Delta = 5$ . To calculate the MED for each dose-response model as well as to use the model averaging approach, we simply need to run Program 4.8. The MEDs based on the Linear, LogLinear, Emax, and Exponential models are 88.4, 65.1, 65, and 94.1 mg, respectively. It is helpful to compare these MEDs to the doses that were identified in the unadjusted setting (see Table 4.9). It is clear that accounting for the two covariates in the candidate models resulted in lower target doses.

Further, the model-specific weights can be computed based on the AIC values as before, which results in the following weights for the Linear, LogLinear, Emax, and Exponential models: 0.399, 0.320, 0.158, and 0.123. These weights are very close to those listed in Table 4.10, which simply reflects the fact that the covariate adjustment did not influence the relative importance of the four dose-response models. The model-averaged MED is computed by running Program 4.9 and is equal to 78 mg.

### 4.4.3 Summary

This section provided a detailed illustration of dose-response tests and dose-finding strategies based on the MCP-Mod algorithm. We used the Phase II trial in patients with schizophrenia from Case study 1 to define the specific steps and SAS tools to identify optimal contrasts; carry out dose-response tests; perform a multiplicity adjustment to choose the final set of dose-response models; and, finally, support dose selection based on one or more promising models. The SAS macros presented in this section (`%RegularOptimalContrasts` and `%GeneralizedOptimalContrasts`) are very flexible and support simple ANOVA-type models as well as more general models that incorporate important covariates. The macros can be applied to dose-finding trials with general unbalanced designs and unequally spaced dose levels. It will be shown in Section 4.5 how these tools can be used to perform dose-response analysis in clinical trials with non-normally distributed endpoints.

From a practical perspective, it is helpful to keep the following points in mind when applying the MCP-Mod procedure in dose-finding trials:

- When selecting candidate dose-response models, it is ideal to identify a diverse set of plausible functions in Step 1. If the models are very similar to each other (e.g., the logistic and SigEmax models), the optimal contrasts derived from the models will be strongly correlated with each other, which might affect the calculation of the multiplicity-adjusted critical value in Step 3. In addition, if too many models are included in the candidate set, the spread of model weights increases, which ultimately impacts the performance of the model averaging approach, which is aimed at selecting a single minimum effective dose based on several dose-response functions in Step 5.
- PROC NLMIXED and PROC NLIN both model nonlinear dose-responses. PROC NLIN estimates parameters using least squares, while PROC NLMIXED relies on maximum likelihood estimates and supports nonlinear models with normally distributed random effects. PROC NLMIXED is recommended as a versatile procedure that streamlines the process of fitting a broad class of nonlinear dose-response models in Step 4. When fitting nonlinear models, it is important to choose meaningful initial values for the model parameters. The initial values must be reasonably justified from the fitted models rather than arbitrarily chosen. Constraints on the initial values should be minimal (as specified by the BOUNDS statement in PROC NLMIXED) as a model should be allowed to scan the parameter space. Artificial constraints can lead to local maxima and potentially spurious results.
- Several approaches are available to support MED selection based on relevant information criteria (e.g., AIC or BIC), maximum test statistic, and model averaging in Step 5. These methods are likely to produce a range of doses, and, if the resulting dose range is too wide, it is helpful to incorporate safety considerations and other clinical information to arrive at the recommended dose.

---

## 4.5 Dose finding in Case study 2

Case study 1 was used in Section 4.4 to illustrate dose finding in a relatively straightforward setting. Two important assumptions made in Section 4.4 include, first, the assumption of a single outcome per patient and, second, the assumption that all outcomes are observed. In fact, simplistic LOCF-based imputation was applied to create a complete data set and satisfy the second assumption in Case study 1. This section extends the MCP-Mod dose-finding algorithm to a setting with

longitudinal observations and missing values. The `%GeneralizedOptimalContrasts` macro will be used in this section to identify optimal dose-response contrasts in trials with unbalanced designs and unequally spaced dose levels.

### 4.5.1 MCP-Mod procedure based on repeated-measures models

As a starting point, the MCP-Mod procedure will be applied to build dose-response models with longitudinal observations in the urticaria clinical trial from Case study 2 by ignoring the missed visits. The more challenging task of dose finding in trials with missing data will be considered in Section 4.5.2.

#### Step 1: Candidate models

The MCP-Mod procedure will be applied to characterize the dose-response relationship and identify target doses in the urticaria clinical trial, based on three candidate models defined in Table 4.4: the Linear, SigEmax, and Exponential models. Recall that the primary endpoint is binary in this trial (response based on the Hives Severity score), and the principles of generalized estimating equations (GEE) will be applied to develop a generalized MCP-Mod solution. The key idea used in this solution is to separate the dose-response model from the expected response and focus on the more general characteristics of the response distribution (Pinheiro, Bretz, and Bornkamp, 2014; Mercier et al., 2015). As shown below, the MCP-Mod dose-finding algorithm can be implemented in the urticaria clinical trial using the same steps and same general approaches that were successfully used in Case study 1.

The general set up for the dose-finding algorithm in this case study is provided by Program 4.14. This program lists the selected candidate models as well as the dose levels and sample sizes in each trial arm. To complete the specification of the candidate models, we need, as before, to discuss initial values of the model parameters. The model based on a linear fit does not require any start values, but the initial parameter values need to be specified in the other two models. For the SigEmax model, the pairs of the dose and expected probability values are guessed based on the maximum treatment effect, i.e.,  $E_{\max}$ . Historical information from similar clinical trials suggests that 60% of  $E_{\max}$  is likely to be achieved at 0.75 mg and 90% at 0.9 mg. Considering the Exponential model, it is assumed that the dose of 0.75 mg is expected to have an increase over placebo with a rate of 60%. These assumptions are passed to the `%SigEmax` and `%Exponential` macros to find the initial values of the model parameters.

#### PROGRAM 4.14 Candidate models in Case study 2

```

data models;
input models $20. ;
datalines;
    Linear
    SigEmax
    Exponential
run;

data dose;
input dose n;
datalines;
    0.00 30
    0.50 30
    0.75 30

```

```

0.90 30
1.00 30
run;

%let delta=1;
%let offset=0.1;
%let MaxEff=1;
%let scaling_factor=1.2;

%StartUp;

%SigEmax(d1=0.75, p1=0.6, d2=0.9, p2=0.9);
%Exponential(d=0.75, p=0.6);

```

## Step 2: Optimal contrasts

Program 4.15 derives the optimal contrasts associated with the three candidate models using the initial estimates and assumptions stored in macro variables created by the %StartUp macro in Program 4.14. The optimal contrast coefficients are computed by calling the %GeneralizedOptimalContrasts macro.

### PROGRAM 4.15 Optimal contrasts in Case study 2

```

proc genmod data=cs2 order=data;
  class visit dose subjid response;
  model response=dose/link=logit dist=bin noint;
    estimate "placebo" dose 1 0 0 0 0 /exp;
    estimate "0.5 mg" dose 0 1 0 0 0 /exp;
    estimate "0.75 mg" dose 0 0 1 0 0 /exp;
    estimate "0.9 mg" dose 0 0 0 1 0 /exp;
    estimate "1 mg" dose 0 0 0 0 1/exp;
    repeated subject=subjid/type=ind covb;
    ods output GEENCov=S(drop=RowName)
      GEEEmpPEst=muHat(where=(Parm ne 'Intercept'));
  run;

%GeneralizedOptimalContrasts;

```

The optimal contrasts computed by Program 4.15 are shown in Table 4.15. In addition, the correlation matrix of the optimal contrasts is shown in Table 4.16. The three resulting contrasts are very strongly correlated with all correlation coefficients exceeding 90%.

**TABLE 4.15** Model-specific optimal contrast coefficients

Dose	Placebo	0.5 mg	0.75 mg	0.9 mg	1 mg
Linear	-0.802	-0.144	0.175	0.320	0.451
SigEmax	-0.559	-0.494	0.135	0.418	0.500
Exponential	-0.709	-0.264	0.085	0.330	0.558

**TABLE 4.16** Correlations among optimal contrasts

	Linear	SigEmax	Exponential
Linear	1.000	0.903	0.979
SigEmax	0.903	1.000	0.956
Exponential	0.979	0.956	1.000

### Step 3: Dose-response tests

To assess the strength of dose-response trends based on the three contrasts identified in Step 2, Program 4.16 uses PROC MIXED to fit repeated-measures models and compute the test statistic for each contrast.

#### PROGRAM 4.16 Dose-response tests based on optimal contrasts in Case study 2

```
proc sql;
    select Linear into :linear_cont separated by " "
        from optcont;
    select SigEmax into :sigemax_cont separated by " "
        from optcont;
    select Exponential into :exponential_cont separated by " "
        from optcont;

proc mixed data=cs2 order=data method=reml;
    class dose visit response;
    model response=dose;
    repeated visit/subject=subjid;
    estimate 'Linear'      dose &linear_cont;
    estimate 'SigEmax'     dose &sigemax_cont;
    estimate 'Exponential' dose &exponential_cont;
run;
```

The resulting test statistics are displayed in Table 4.17. The trend test based on the SigEmax contrast is non-significant even before a multiplicity adjustment. Note that the raw two-sided  $p$ -value is greater than 0.05, and, with any multiplicity adjustment, it will only increase. It is clear that this model will be excluded in the final set of clinically relevant models.

**TABLE 4.17** Output from Program 4.16

Model	Contrast estimate	Standard error	Test statistic	P-value
Linear	0.1224	0.0512	2.35	0.0199
SigEmax	0.0978	0.0519	1.89	0.0614
Exponential	0.1218	0.0519	2.35	0.0204

To determine whether the other two dose-response contrasts are statistically significant in the context of this multiplicity problem, the %CriticalValue macro is invoked. The first parameter of the macro is the data set with the correlation matrix computed in Program 4.15. To obtain the number of degrees of freedom (i.e.,  $\nu = N - m$ ), recall that the total number of data points in the trial is  $N = 450$  (150 patients with three post-baseline visits), but the primary outcome is missing at eight visits. Also, the total number of trial arms is  $m = 5$ ; and thus,  $\nu = 450 - 8 - 5 = 437$ . Program 4.17 executes the %CriticalValue macro to derive the multiplicity-adjusted critical value in this problem.

#### PROGRAM 4.17 Multiplicity-adjusted critical value in Case study 2

```
%CriticalValue(corr=corrmat, nu=437, alpha=0.025);
```

The adjusted critical value is 2.130, and it follows from Table 4.17 that the contrasts derived from the Linear and Exponential models result in a significant dose-response relationship after the adjustment.

## Step 4: Model selection

It is shown in Step 3 that the Linear and Exponential models are both clinically relevant, and Program 4.18 uses PROC NLMIXED to fit these models and compute the parameter estimates as well as the values of the information criterion (AIC).

### PROGRAM 4.18 Parameter estimates of the significant covariate-adjusted model in Case study 2

```
* Linear model;
proc nlmixed data=cs2;
  parms s2=0.5;
  eta=e0+delta*dose;
  y=(eta+b)*time;
  p = exp(y)/(1 + exp(y));
  model response~binomial(1, p);
  random b~normal(0,s2) subject=subjid;
  ods output ParameterEstimates=LinParameterEstimates
    FitStatistics=AIC1(rename=(Value=AIC)
      where=(Descr='AIC (smaller is better)' ));
  predict e0+delta*dose out=LinPrediction;
run;

* Exponential model;
proc nlmixed data=cs2;
bounds e1<=1;
  eta=e0+e1*(exp(dose/delta)-1);
  y=(eta+b)*time;
  p = exp(y)/(1+exp(y));
  model response~binomial(1, p);
  random b~normal(0,s2) subject=subjid;
  ods output ParameterEstimates=ExpParameterEstimates
    FitStatistics=AIC2(rename=(Value=AIC)
      where=(Descr='AIC (smaller is better)' ));
  predict e0+e1*(exp(dose/delta)-1) out=ExpPrediction;
run;

data aic;
  set aic1 aic2;
run;
```

Table 4.18 lists the model parameters and model-specific AIC values. The two test statistics are very close to each other, and so are the AIC values. In this case, it might be prudent to consider MED estimation using the model averaging technique rather than focusing on the MEDs derived from each model.

**TABLE 4.18 Dose-response models based on optimal contrasts**

Model	Parameter estimates	Test statistic	AIC
Linear	$E_0 = -0.015, \delta = 0.66$	2.35	521.6
Exponential	$E_0 = 0.008, E_1 = 1.0, \delta = 1.91$	2.35	521.3

## Step 5: Target dose selection

The primary endpoint in Case study 2 is binary, and the clinically relevant treatment effect, which is the key component of estimating the minimum effective dose, can be

defined in terms of the response rates or on a logit scale. It will be assumed that a 25% relative improvement over the placebo response rate is a clinically meaningful treatment difference. If the response rate in the placebo arm is 50%, the MED will be defined as the dose that results in the response rate of 62.5%. Using a logit scale, the logit of the placebo response rate is  $\text{logit}(0.5) = 0$ , and the logit of the response rate that corresponds to the desirable effect is  $\text{logit}(0.625) = 0.5$ . Using the difference between the two values, the clinically acceptable improvement on a logit scale is set to  $\Delta = 0.5$ .

To illustrate the process of estimating the MED in this case study based on the two models in the final set, PROC NLMIXED from Program 4.18 predicts the response rate and its logit in the five trial arms (the results are saved in the **LinPrediction** and **ExpPrediction** data sets). The predicted values are shown in Table 4.19. We can see from this table that, with either dose-response model, a 25% relative improvement over placebo is achieved between 0.75 mg and 0.9 mg. The minimum effective doses are found using straightforward inverse regression techniques and are shown in Table 4.20.

**TABLE 4.19** Predicted response rates

Dose (mg)	Linear model		Exponential model	
	Logit scale	Response rate	Logit scale	Response rate
0	-0.015	0.496	0.008	0.502
0.5	0.315	0.578	0.308	0.576
0.75	0.480	0.618	0.490	0.620
0.9	0.579	0.641	0.611	0.648
1.0	0.645	0.656	0.697	0.668

**TABLE 4.20** Model-specific minimum effective doses

Model	MED (mg)
Linear	0.76
Exponential	0.77

Since the model-specific MEDs are virtually identical, we should expect that the weighted target dose based on model averaging will also be close to 0.77 mg. Indeed, assuming a non-informative prior with  $p_1 = p_2 = 1/2$ , the model-specific weights based on the AIC are computed using the values of this information criterion shown in Table 4.18. It is easy to check that  $w_1 = 0.461$  and  $w_2 = 0.539$ . The minimum effective dose based on the weighted average of the two models is computed as follows:

$$w_1 \text{MED}_1 + w_2 \text{MED}_2 = 0.77.$$

Thus, model averaging based on the AIC suggests 0.77 mg as the optimal dose in Case study 2.

#### 4.5.2 MCP-Mod procedure in the presence of missing observations

It was stressed at the beginning of Section 4.5.1 that the missing visits in the urticaria clinical trial would simply be ignored when dose finding is performed. However, ignoring missing observations is likely to introduce bias since the observations might be attributed to the experimental treatment's lack of efficacy. Under such circumstances, it is recommended to perform sensitivity assessments by applying the

MCP-Mod algorithm under several relevant imputation strategies. In this section, we compare and contrast three different imputation strategies. The first two strategies focus on “stress testing” the algorithm by considering simple imputation techniques that make conservative and liberal assumptions about the missing observations. The third strategy also entertains the idea of “conservative imputation” but employs a more sophisticated and statistically principled approach that relies on multiple imputation to account for uncertainty due to missing data. For more information on commonly used methods for handling missing data in clinical trials, see Chapter 7, O’Kelly and Ratitch (2014), and Mallinckrodt and Lipkovich (2016, Section III).

## Conservative and liberal imputation strategies

The conservative imputation strategy relies on the assumption that the missed visits in the trial were caused by lack of efficacy. To implement the strategy, the missing outcomes are set to 0 in the `cs2` data set from the urticaria trial, i.e., it is assumed that the corresponding patients failed to respond to treatment (see Program 4.19). This imputation strategy, known as *non-responder imputation*, needs to be applied uniformly across the trial arms to avoid bias.

Program 4.19 executes the `%GeneralizedOptimalContrasts` macro to find the optimal dose-response contrasts derived from the same three candidate model as in Section 4.5.1 (i.e., Linear, SigEmax, and Exponential models). The imputed data set (`cs2imp1` data set) is used in this analysis.

### PROGRAM 4.19 Dose-finding based on non-responder imputation in Case study 2

```

data cs2imp1;
  set cs2;
  if missing(response) then response=0;
run;

proc genmod data=cs2imp1 order=data;
  class visit dose subjid response;
  model response=dose/link=logit dist=bin noint;
  repeated subject=subjid/type=ind covb;
  ods output GEEVCov=S(drop=RowName)
    GEEEmpPEst=muHat(where=(Parm ne 'Intercept'));
run;

%GeneralizedOptimalContrasts;

```

Program 4.19 generates the optimal contrasts shown in Table 4.21. The dose-response tests based on these contrasts are carried out in the usual manner (see Program 4.16), and the `%CriticalValue` macro is called as in Program 4.17 to compute the multiplicity-adjusted critical value that accounts for the joint distribution of the test statistics. It is important to note that the number of degrees of freedom in the macro call is  $\nu = 445$ , since there are no missing observations in the `cs2imp1` data set. It can be shown that, as in Section 4.5.1, only two models (Linear and Exponential models) exhibit a significant dose-response after the multiplicity

**TABLE 4.21** Optimal contrasts based on non-responder imputation

Dose	Placebo	0.5 mg	0.75 mg	0.9 mg	1 mg
Linear	-0.802	-0.144	0.171	0.326	0.449
SigEmax	-0.557	-0.496	0.132	0.425	0.496
Exponential	-0.709	-0.265	0.084	0.336	0.554

adjustment. The parameter estimates in these two models and associated values of the AIC are displayed in Table 4.22.

**TABLE 4.22** Dose-response models based on non-responder imputation

Model	Parameter estimates	Test statistic	AIC
Linear	$E_0 = -0.115, \delta = 0.715$	2.52	538.7
Exponential	$E_0 = -0.100, E_1 = 1.0, \delta = 1.783$	2.49	538.6

The minimum effective doses based on the Linear and Exponential models are 0.7 mg and 0.72 mg, respectively. Using the model averaging approach based on the AIC suggests an MED of 0.71 mg for the conservative imputation strategy. The resulting target doses are slightly lower than those that were estimated in Section 4.5.1 when the missed visits were ignored in the dose-response analysis.

The non-responder imputation technique relies on a very conservative assumption that the missed visits correspond to lack of efficacy. In a similar way, we can consider the other extreme where the patients who missed visits are assumed to be responders. This imputation technique, known as *responder imputation*, is easy to implement by replacing each missing outcome with 1 in the *cs2* data set. Program 4.20 creates the imputed data set (*cs2imp2* data set), and the MCP-Mod algorithm can be applied to this data set to compute the optimal contrasts, select clinically relevant models, and identify the target dose (minimum effective dose).

#### **PROGRAM 4.20 Dose-finding based on responder imputation in Case study 2**

```
data cs2imp2;
  set cs2;
  if missing(response) then response=1;
run;
```

It can be shown that the Linear and Exponential models are included in the final set, and the MEDs computed from the two dose-response models are 0.81 mg and 0.83 mg, respectively. The model averaging approach using AIC-based weights suggests an MED of 0.82 mg for the strategy that relies on responder imputation. With this liberal imputation strategy, the MEDs are slightly higher than the doses based on the strategy that ignored the missed visits (see Section 4.5.1).

#### **Pattern-mixture imputation strategy with placebo imputation**

As an alternative to the very conservative and very liberal imputation techniques described above, we will consider a more robust approach in which the missing outcomes are assumed to resemble the effect in the placebo arm. This approach is an example of *pattern-mixture imputation*, which, in this particular case, uses placebo imputation. Several strategies within this general approach are available and include imputation of all missed visits, imputation of selected missed visits, and imputing the missed visits after the last non-missed visit. For more information on these and related strategies, see, for example, O'Kelly and Ratitch (2014).

Here we provide a rather intuitive and exploratory approach to multiple imputation in the context of evaluating optimal doses and MED estimating using the MCP-Mod methodology. The approach can be outlined as follows:

- Generate 100 completed data sets for placebo patients with outcomes imputed using methods that rely on the MAR (missingness at random) assumption. These data sets will be used as the basis for constructing placebo-based imputations in the treatment arm.

- Generate 100 copies of the data set (samples) for the treatment arm with missing values. For each sample:
  - Complete the data by imputing missing outcomes using an imputation model estimated from a corresponding completed placebo data set.
  - Apply the MCP-Mod algorithm to identify the optimal dose-response model and corresponding MED estimates, including estimates based on model averaging.
- Assess (informally) the estimated dose-response model parameters and the consistency of MEDs across the 100 imputed data sets.

In the following, we provide details and SAS code for implementing the above steps. It is worth noting that, while in our implementation we used relatively simple imputation models that only take into consideration baseline covariates, more sophisticated models using all observed data on longitudinal profiles for each patient could be considered. Also, more formal procedures for combining optimal contrasts associated with different dose-response models across the samples could have been developed and bootstrap-based multiplicity-adjusted confidence intervals for optimal doses could have been constructed.

Using the general set up introduced at the beginning of Section 4.5.1, pattern-mixture modeling will be performed as follows. Before applying the MCP-Mod dose-finding algorithm, the `cs2` data set needs to be “prepped” for placebo imputation. The key to a successful imputation procedure is ensuring that the imputation strategy captures all the inherent variability in the data. To accomplish this, several imputations are required and, to select the number of imputations, multiple runs might be needed until the results stabilize. In this case, for the sake of illustration, 100 imputations were performed.

Program 4.21 defines the first step in this imputation strategy. Recall from Table 4.1 that several patients in the placebo arm had missed visits. To account for this, the placebo outcomes are imputed first as seen in this program. Since the primary endpoint is binary, a monotone logistic model is appropriate to impute the missing placebo outcomes. The number of imputed data sets generated from the `cs2` data set is determined by the `nimpute` macro variable, which is set to 100. The results are saved in the `cs2placebo` data set.

#### **PROGRAM 4.21    Multiple imputation to construct complete data sets in the placebo arm in Case study 2**

```
%global nimpute;
%let nimpute=100;

proc mi data=cs2 out=cs2placebo nimpute=&nimpute. seed=123;
  var gender response;
  class gender response;
  where dose=0;
  monotone logistic;
run;
```

The next step in Program 4.22 duplicates the outcomes in the active treatment arms and combines them with the imputed placebo outcomes. It is assumed that the missing outcomes in the active treatment arms are missing not at random (MNAR). Sensitivity analyses based on the assumptions of missing at random and missing completely at random are suggested. However, the selection criteria should be based on statistical rationale. The imputed data set (`cs2imp3` data set) is created so that there are no missed visits in the active arms since the missed outcomes are modeled based on the corresponding placebo outcome using an imputation model that incorporates an important covariate (gender). Note that the NIMPUTE option

in PROC MI limits the imputation to a single placebo imputation in each of the 100 copies. Thus, Program 4.22 creates 100 unique copies of the `cs2` data where there are no missing outcomes. These data sets capture the variability of the missing outcomes in the active treatment arms that are consistent with the effect in the placebo arm.

### **PROGRAM 4.22 Pattern-mixture using placebo Imputation in Case study 2**

```

data cs2dose;
set cs2(where=(dose ne 0));
do _imputation_=1 to &nimpute. ;
  output;
end;
run;

data cs2all;
  set cs2placebo cs2dose;
run;

proc sort data=cs2all;
  by _imputation_ dose subjid visit;
run;

proc mi data=cs2all out=cs2imp3 nimpute=1 seed=123;
  by _imputation_;
  var dose gender response;
  class gender response dose;
  mnar model(response/modelobs=(dose="0"));
  monotone logistic(response=gender/descending);
run;

```

The results generated by Program 4.22 are stored in the `cs2imp3` data set. Table 4.23 provides a summary of the 100 imputed data sets. Note that there are no missing outcomes in any of the trial arms.

Typically, PROC MI is followed by PROC MIANALYZE where the point estimates and associated variance-covariance matrices, obtained by applying complete-data analysis to each imputed set, are pooled using Rubin's combination rules. Recall that each imputed data set is analyzed using the MCP-Mod algorithm, and the best dose is determined in each data set based on the optimal dose-response curve selected from all the candidate models. However, incorporating the model selection step within each imputed data set presents an issue, because Rubin's estimator assumes that each data set is analyzed using the same model and does not account for the model selection step. Therefore, model selection should be performed *after* pooling the estimates across the imputed data sets. Computing the pooled point estimates for the optimal contrasts associated with each dose-response curve can be easily done by averaging the contrasts from the individual data sets. However, computing the corresponding "pooled" critical values can be challenging. This creates a cumbersome problem of combining correlation matrices (computed for the covariate-adjusted test statistics under the null hypothesis of no dose effect) over the 100 imputed data sets, which is needed for computing a single set of multiplicity-adjusted critical values for candidate dose-response models. Note also that Rubin's combined inference assumes MAR, and here the imputation is effectively performed under MNAR (missingness not at random), as a result of "borrowing" imputation models from the placebo arm. Since we are limited in the usage of Rubin's rules, we incorporate a separate assessment of each imputed data set. Each data set will be analyzed using the MCP-Mod algorithm presented in Section 4.5.1 as described below.

**TABLE 4.23 Summary of outcomes by trial arm and visit in the 100 imputed data sets**

Imputation	Trial arm	Outcome	Visit 1	Visit 2	Visit 3
1	Placebo	No response	12	18	16
		Response	18	12	14
	0.5 mg	No response	12	7	15
		Response	18	23	15
	0.75 mg	No response	13	15	14
		Response	17	15	16
	0.9 mg	No response	10	7	11
		Response	20	23	19
	1 mg	No response	9	11	11
		Response	21	19	19
2	Placebo	No response	12	18	14
		Response	18	12	16
	0.5 mg	No response	12	7	14
		Response	18	23	16
	0.75 mg	No response	13	15	14
		Response	17	15	16
	0.9 mg	No response	10	9	11
		Response	20	21	19
	1 mg	No response	9	11	11
		Response	21	19	19
...					
100	Placebo	No response	13	18	14
		Response	17	12	16
	0.5 mg	No response	12	7	14
		Response	18	23	16
	0.75 mg	No response	13	15	14
		Response	17	15	16
	0.9 mg	No response	10	8	11
		Response	20	22	19
	1 mg	No response	9	11	11
		Response	21	19	19

## Step 1: Candidate models

This step is identical to Step 1 of the MCP-Mod algorithm from Section 4.5.1, and the same three candidate models are chosen (Linear, SigEmax, and Exponential models).

## Step 2: Optimal contrasts

The initial parameter estimates are stored in appropriate macro variables, and a generalized solution is obtained for each imputation as shown in Program 4.23. This results in 100 sets of optimal contrasts and corresponding correlation matrices, one for each imputed data set from Program 4.23. The 100 sets of optimal contrasts are stored in the `OptContr` data set, and the 100 correlation matrices are stored in the `CorrelationMat` data set.

### PROGRAM 4.23 Optimal contrasts in the 100 imputed data sets in Case study 2

```
proc genmod data=cs2imp3 descend order=data;
  by _Imputation_;
  class visit dose subjid;
  model response=dose/link=logit dist=bin noint;
  repeated subject=subjid/type=ind covb ;
```

```

ods output GEECov=cov(drop=rownname)
      GEEEmpPEst=mu(where=(Parm ne 'Intercept'));
run;

* Optimal contrast and correlation matrix for each imputed data set;
%macro ImputedGenOptCont(nimpute=);
%do p=1 %to &nimpute;
  Data S(drop=_Imputation_);
  Set Cov;
  where _Imputation_=&p.;
run;

Data muHat(drop=_Imputation_);
Set mu;
where _Imputation_=&p.;
run;

%GeneralizedOptimalContrasts;

data Opt_Contr_&p.;
set optcont;
_imputation_=&p.;
run;

data Correlation_Mat_&p.;
set corrrmat;
_imputation_=&p.;
run;
%end;
%mend ImputedGenOptCont;

%ImputedGenOptCont(nimpute=&nimpute.);

data OptContr;
  set Opt_Contr_1 - Opt_Contr_&nimpute.;
run;

data CorrelationMat;
  set Correlation_Mat_1 - Correlation_Mat_&nimpute.;
run;

```

The resulting sets of optimal contrasts are presented in Table 4.24. It is helpful to note that, due to the small number of missed visits in this large data set, the imputation-specific correlation matrices are all equal to each other when rounded to the third decimal.

**TABLE 4.24 Model-specific optimal contrast coefficients in the 100 imputed data sets**

Imputation	Model	Placebo	0.5 mg	0.75 mg	0.9 mg	1 mg
1	Linear	-0.802	-0.142	0.174	0.318	0.452
	SigEmax	-0.558	-0.495	0.136	0.415	0.502
	Exponential	-0.710	-0.264	0.087	0.328	0.559
2	Linear	-0.802	-0.142	0.171	0.326	0.448
	SigEmax	-0.560	-0.492	0.131	0.425	0.496
	Exponential	-0.710	-0.262	0.083	0.336	0.554
...						
100	Linear	-0.803	-0.141	0.172	0.322	0.450
	SigEmax	-0.561	-0.492	0.133	0.420	0.499
	Exponential	-0.711	-0.262	0.085	0.332	0.556

### Step 3: Dose-response tests

To carry out the dose-response tests based on the optimal contrasts in the imputed data sets, PROC MIXED is used in Program 4.24 to fit the candidate models.

**PROGRAM 4.24 Dose-response tests based on optimal contrasts in the 100 imputed data sets in Case study 2**

```
%macro SignificantFit;
%do p=1 %to &nimpute;
proc sql;
select linear into :linear_contr separated by " "
from optcontr where _imputation_=&p.;
select sigemax into :sigemax_contr separated by " "
from optcontr where _imputation_=&p.;
select exponential into :exponential_contr separated by " "
from optcontr where _imputation_=&p.;

proc mixed data=cs2imp3 order=data method=reml;
where _imputation_=&p.;
class dose visit response;
model response=dose;
repeated visit/subject=subjid;
Estimate 'Linear'      dose &Linear_Contr.;
Estimate 'SigEmax'     dose &SigEmax_Contr.;
Estimate 'Exponential' dose &Exponential_Contr.;
ods output estimates=estimates;
run;

data estimates_&p.;
set estimates;
_imputation_=&p.;
run;
%end;
%mend SignificantFit;

%SignificantFit;

data estimates;
set estimates_1 - estimates_&nimpute.;
run;
```

The contrast-specific test statistics displayed in Table 4.25 and produced by Program 4.24 show a certain amount of variability across the imputed data sets.

**TABLE 4.25 Model-specific test statistics in the 100 imputed data sets**

Imputation	Model	Contrast estimate	Test statistic
1	Linear	0.1278	2.49
	SigEmax	0.1067	2.08
	Exponential	0.1274	2.48
2	Linear	0.1018	1.98
	SigEmax	0.0806	1.57
	Exponential	0.1020	1.98
...			
100	Linear	0.1141	2.22
	SigEmax	0.0912	1.78
	Exponential	0.1134	2.21

But they tend to be fairly consistent in the sense that the contrasts derived from the SigEmax model are clearly less significant than the other two contrasts.

The multiplicity-adjusted critical value for each imputed data set is derived in Program 4.25. The program invokes the %CriticalValue macro with the number of degrees of freedom set to  $\nu = N - m = 450 - 5 = 445$ .

#### **PROGRAM 4.25 Multiplicity-adjusted critical values in the 100 imputed data sets in Case study 2**

```
%macro CriticalValueLoop;
%do p=1 %to &nimpute;
  data correlation_mat(drop=_imputation_);
    set correlationmat;
    where _imputation_=&p.;
  run;

  %CriticalValue(corr=correlation_mat, nu=445, alpha=0.025)

  data critvalue_&p.;
    set critvalue;
    _imputation_=&p.;
  run;
%end;
%mend CriticalValueLoop;

%CriticalValueLoop;

data critvalue;
  set critvalue_1 - critvalue_&nimpute.;
run;

data estimates_cv;
merge estimates critvalue;
  by _imputation_;
  if tvalue<critvalue then delete;
run;
```

Each multiplicity-adjusted critical value is based on the correlation matrix of each imputed data set and different critical values are generally expected. However, as pointed out above, due to the reasonably similar correlation matrices, the adjusted critical values are identical in this clinical trial example. The common adjusted critical value is 2.11, and, based on this value, the SigEmax contrast is not significant in any of the imputations. The contrasts derived from the Linear and Exponential models are significant at the adjusted level in all imputed data sets. Hence, the two models will be included in the final set of clinically relevant models.

#### **Step 4: Model selection**

Program 4.26 uses PROC NL MIXED to fit the dose-response models and compute the parameter estimates as well as the value of the information criterion (AIC) for each imputed data set.

#### **PROGRAM 4.26 Parameter estimation in clinically relevant models in the 100 imputed data sets in Case study 2**

```
proc sort data=cs2imp3;
  by _imputation_ dose subjid;
run;
```

```

* Linear model;
proc nlmixed data=cs2imp3;
    by _imputation_;
    parms s2=0.05;
    eta=e0+delta*dose;
    y=(eta+b)*time;
    p = exp(y)/(1 + exp(y));
    model response ~ binomial(1, p);
    random b ~ normal(0,s2) subject=subjid;
    ods output parameterestimates=linparameterestimates
        fitstatistics=aic1(rename=(value=aic_lin)
        where=(Descr='AIC (smaller is better)' ));
run;

* Exponential model;
proc nlmixed data=cs2imp3;
by _imputation_;
bounds e1 <= 1;
eta=e0+e1*(exp(treatment/delta)-1);
y=(eta+b)*time;
p = exp(y)/(1 + exp(y));
model response ~ binomial(1, p);
random b ~ normal(0,s2) subject=subjid;
ods output parameterestimates=expparameterestimates
    fitstatistics=aic2(rename=(value=aic_exp)
    where=(Descr='AIC (smaller is better)' ));
run;

data aic;
    set aic1 aic2;
run;

```

The parameter estimates and AIC values computed by Program 4.26 are summarized in Table 4.26. A careful review of the AIC values in this table reveals that the Linear model tends to provide a similar or better fit compared to the Exponential model across the imputed data sets.

**TABLE 4.26 Dose-response models in the 100 imputed data sets**

Imputation	Model	Parameter estimates	Test statistic	AIC
1	Linear	$E_0 = -0.095, \delta = 0.685$	2.49	545.0
	Exponential	$E_0 = 0.078, E_1 = 0.652, \delta = 1.18$	2.48	547.4
2	Linear	$E_0 = 0.027, \delta = 0.558$	2.22	538.6
	Exponential	$E_0 = 0.069, E_1 = 0.671, \delta = 1.13$	2.21	538.7
...				
100	Linear	$E_0 = 0.000, \delta = 0.619$	2.22	534.7
	Exponential	$E_0 = 0.058, E_1 = 0.678, \delta = 1.12$	2.21	534.9

### Step 5: Target dose selection

The last step in the dose-finding algorithm focuses on the identification of the minimum effective dose in the urticaria trial based on the clinically relevant threshold. As before, this threshold is defined as a 25% relative improvement over placebo in terms of the response rate or a difference of 0.5 on a logit scale. Using this threshold, Program 4.27 estimates the MED in each imputed data set.

**PROGRAM 4.27 MED in Case study 2**

```

data linparameterestimates_;
merge linparameterestimates(where=(parameter='delta')) aic1;
  by _imputation_;
    med_lin=0.5/estimate;
run;

data display1;
set linparameterestimates_;
keep _Imputation_
run;

proc transpose data=expparameterestimates out=expparameterestimates_;
var estimate;
  id parameter;
  by _imputation_;
run;

data expparameterestimates_;
merge expparameterestimates_ aic2;
  by _imputation_;
    med_exp=delta*log(1+0.5/e1);
run;

data med;
merge linparameterestimates_ expparameterestimates_;
  by _imputation_;
  keep aic_lin med_lin aic_exp med_exp _imputation_;
run;

data med;
set med;
  by _imputation_;
  p=1/2;
  array x[*] aic_lin aic_exp;
  array a[*] a1-a2;
  array b[*] b1-b2;
  do i=1 to 2;
    a{i}=x{i}-x{1};
    b{i}=x{i}-x{2};
  end;
  wt1=p/(p*(exp(-a1/2)+exp(-a2/2)));
  wt2=p/(p*(exp(-b1/2)+exp(-b2/2)));
  med=wt1*med_lin+wt2*med_exp;
run;

proc means data=med mean;
  var med;
run;

```

The MED is computed for each dose-response model from each imputation data set. It is reasonable to use the standard model averaging approach to arrive at a single MED estimate. After that, the final MED estimate can be defined as the average dose over the 100 imputation data sets. The resulting average MED is equal to 0.79 mg.

Based on the three imputation strategies presented in this section, we see that the MED estimates range from 0.72 mg (conservative imputation strategy) and 0.83 mg (liberal imputation strategy). Reviewing the results from the original analysis

from Section 4.5.1 that ignored the missing observations, the MED was 0.77 mg. These results are fairly consistent and provide useful information to support the clinical decision making aimed at determining the minimum effective dose to be evaluated at the confirmatory stage of this development program.

### 4.5.3 Summary

This section presented an extension of the original MCP-Mod dose-finding algorithm to a general setting with non-normally distributed endpoints. The algorithm is easily extended to Phase II clinical trials with binary, count, and time-to-event outcomes. In addition, as shown in this section, standard sensitivity assessments used in clinical trials with missing observations can be applied to perform similar sensitivity analyses in the context of dose finding. This includes simple imputation techniques such as responder and non-responder techniques as well as more sophisticated model-based imputation methods (e.g., pattern-mixture imputation). Application of several imputation methods results in a range of target doses that represent possible decisions under different assumptions about the missing outcomes. As emphasized throughout the chapter, a range of doses identified by the MCP-Mod algorithm in settings with a complete data set or in settings with missing observations provides a useful guideline that needs to be used along with clinical judgment and other relevant considerations to support the dose selection decisions for confirmatory trials.

## 4.6 References

---

- Abelson, R.P., Tukey, J.W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics* 34, 1347-1369.
- Bornkamp, B., Bretz, F., Dmitrienko, A., Enas, G., Gaydos, B., Hsu, C.H., König, F., Krambs, M., Liu, Q., Neuenschwander, B., Parke, T., Pinheiro, J., Roy, A., Sax, R., Shen, F. (2007). Innovative approaches for designing and analyzing adaptive dose-ranging trials. *Journal of Biopharmaceutical Statistics* 17, 965-995.
- Bornkamp, B., Pinheiro, J., Bretz, F. (2009). MCPMod: An R package for the design and analysis of dose-finding studies. *Journal of Statistical Software* 29, 1-23.
- Bornkamp, B., Bretz, F., Dette, H., Pinheiro, J. (2011). Response-adaptive dose-finding under model uncertainty . *The Annals of Applied Statistics* 5, 1611-1631.
- Bretz, F., Pinheiro, J.C., Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* 61, 738-748.
- Bretz, F., Tamhane, A.C., Pinheiro, J. (2009). Multiple testing in dose response problems. *Multiple Testing Problems in Pharmaceutical Statistics*. Dmitrienko, A., Tamhane, A.C., Bretz, F. (editors). New York: Chapman and Hall/CRC Press.
- Buckland, S.T., Burnham, K.P., Augustin, N.H. (1997). Model selection: an integral part of inference. *Biometrics* 53, 275-290.
- Claeskens, G. Hjorth, N.L. (2008). *Model Selection and Model Averaging*. Cambridge, UK: Cambridge University Press.
- Chuang-Stein, C., Agresti, A. (1997). A review of tests for detecting a monotone dose-response relationship with ordinal response data. *Statistics in Medicine* 16, 2599-2618.

- Cross, J., Lee, H., Westelinck, A., Nelson, J., Grudzinskas, C., Peck, C. (2002). Postmarketing drug dosage changes of 499 FDA-approved new molecular entities, 1980-1999. *Pharmacoepidemiology and Drug Safety*. 11, 439-446.
- Dmitrienko, A., Fritsch, K., Hsu, J., Ruberg, S. (2007). Design and analysis of dose-ranging clinical studies. *Pharmaceutical Statistics Using SAS: A Practical Guide*. Dmitrienko, A., Chuang-Stein, C., D'Agostino, R. (editors). Cary, NC: SAS Institute Inc.
- EMA (2014). Qualification opinion of MCP-Mod as an efficient statistical methodology for model-based design and analysis of Phase II dose finding studies under model uncertainty. Committee for Medicinal Products for Human Use (CHMP). EMA/CHMP/SAWP/757052/2013.
- Genz, A., Bretz, F. (2002). Methods for the computation of multivariate *t*-probabilities. *Journal of Computational and Graphical Statistics* 11, 950-971.
- Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science* 4, 382-401.
- Kass, R. E., Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* 90, 773-795.
- Klingenberg, B. (2009). Proof of concept and dose estimation with binary responses under model uncertainty. *Statistics in Medicine* 28, 274-292.
- Mallinckrodt, C., Lipkovich, I. (2016). *Analyzing Longitudinal Clinical Trial Data: A Practical Guide*. New York: Chapman and Hall/CRC Press.
- Menon, S., Zink, R. (2016). *Clinical Trials Using SAS: Classical, Adaptive and Bayesian Methods*. Cary, NC: SAS Institute Inc.
- Mercier, F., Bornkamp, B., Ohlssen, D., Wallstroem, E. (2015). Characterization of dose-response for count data using a generalized MCP-Mod approach in an adaptive dose-ranging trial. *Pharmaceutical Statistics* 14, 359-367.
- O'Kelly, M., Ratitch, B. (editors). (2014). *Clinical Trials with Missing Data: A Guide for Practitioners*. New York: Wiley.
- Pinheiro, J. C., Bornkamp, B., Bretz, F. (2006). Design and analysis of dose finding studies combining multiple comparisons and modeling procedures. *Journal of Biopharmaceutical Statistics* 16, 639-656.
- Pinheiro, J.C., Bretz, F., Branson, M. (2006). Analysis of dose-response studies: Modeling approaches. *Dose Finding in Drug Development*. Ting, N. (editor). New York: Springer.
- Pinheiro, J., Bretz, F., Bornkamp, B. (2014). Generalizing the MCP-Mod methodology beyond normal, independent data. *ASA NJ Chapter 35th Annual Spring Symposium*.
- Pinheiro J., Bornkamp B., Glimm E., Bretz F. (2013). Model-based dose finding under model uncertainty using general parametric models. *Statistics in Medicine* 33, 1646-1661.
- Raftery, A.E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*. Marsden, P. V. (editor), 111-195. Cambridge, MA: Blackwell.
- Ruberg, S.J. (1989). Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association* 84, 816-822.
- Ruberg, S.J. (1995a). Dose response studies. Some design considerations. *Journal of Biopharmaceutical Statistics* 5, 1-14.
- Ruberg, S.J. (1995b). Dose response studies. Analysis and interpretation. *Journal of Biopharmaceutical Statistics* 5, 15-42.
- Scheffe, H. (1959). *The Analysis of Variance*. London, UK: Wiley.

- Stewart, W. H., Ruberg, S. J. (2000). Detecting dose response with contrasts. *Statistics in Medicine* 19, 913-921.
- 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.
- Ting, N. (editor) (2006). *Dose Finding in Drug Development*. New York: Springer.
- Tukey, J.W., Ciminera, J.L., Heyse, J.F. (1985). Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics* 41, 295-301.
- Williams, D.A. (1971). A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics* 27, 103-117.
- Williams, D.A. (1972). The comparison of several dose levels with a zero dose control. *Biometrics* 28, 519-531.

# Chapter 5

## Multiplicity Adjustment Methods

Thomas Brechenmacher (QuintilesIMS)

Alex Dmitrienko (Mediana)

5.1	Introduction	179
5.2	Single-step procedures	184
5.3	Procedures with a data-driven hypothesis ordering	189
5.4	Procedures with a prespecified hypothesis ordering	202
5.5	Parametric procedures	212
5.6	Gatekeeping procedures	221
5.7	References	241
5.8	Appendix	244

This chapter discusses statistical strategies for handling multiplicity issues arising in clinical trials with several clinical objectives. It defines commonly used multiplicity adjustments, including basic single-step testing methods, as well as more advanced stepwise methods relying on data-driven or pre-specified hypothesis ordering. The chapter also reviews gatekeeping testing strategies used in more advanced settings—for example, in trials that support simultaneous evaluation of multiple clinical endpoints at several doses of a novel treatment. The multiplicity adjustments introduced in the chapter are illustrated using confirmatory Phase III clinical trials.

### 5.1 Introduction

---

Multiplicity problems arise in virtually all late-stage clinical trials since trial sponsors are often interested in investigating multiple clinical objectives based on the evaluation of several endpoints, multiple dose-control comparisons, assessment of the treatment effect in two or more patient populations, etc. In these cases, the overall success criterion in a trial is defined on the basis of multiple significance tests. It is, therefore, critical to control the probability of incorrectly concluding that the novel treatment is effective (Type I error rate) at a nominal level, e.g., one-sided  $\alpha = 0.025$  or two-sided  $\alpha = 0.05$ .

Example of multiplicity problems encountered in clinical trials are provided below:

1. **Multiple treatment comparisons.** Multiple testing is commonly encountered in clinical trials involving several treatment groups. Examples of such situations

include Phase II trials that are designed to assess the efficacy and safety profile of several doses of an experimental treatment compared to placebo or to an active control. Performing multiple comparisons of different dose levels of an experimental treatment to a control causes multiplicity problems.

2. **Multiple primary endpoints.** Multiplicity is also caused by multiple criteria for assessing the efficacy or safety of an experimental treatment. The multiple criteria are required to accurately characterize various aspects of the expected therapeutic benefits. For example, the efficacy profile of cardiovascular drugs is typically evaluated using multiple outcome variables such as all-cause mortality, nonfatal myocardial infarction, or refractory angina/urgent revascularization.
3. **Multiple secondary endpoints.** It is commonly accepted that multiplicity issues in the primary analysis must be addressed in all confirmatory clinical trials. In modern clinical trials, it is more and more desirable to also make formal claims based on key secondary endpoints. Proper multiplicity adjustment is then also required for the secondary statistical tests to ensure control of the overall Type I error rate. For example, in oncology clinical trials, it is often desirable to characterize the clinical benefit using both progression-free survival and overall survival.
4. **Multiple patient populations.** The primary analysis in a clinical trial can be performed in the general population of patients as well as a prespecified target subgroup. Patients in this subgroup can be, for example, expected to experience greater treatment benefit or more likely to respond compared with the patients in the general population. Multi-population designs of this kind arise in oncology clinical trials and other therapeutic areas.

These clinical trial examples deal with a single *source of multiplicity* or a single design feature that contributes to Type I error rate inflation. It is increasingly more common to encounter problems with several sources of multiplicity in confirmatory clinical trials. These problems are characterized by the fact that multiple features that result in an inflated Type I error rate are considered. Examples include clinical trials that evaluate the effect of several endpoints at multiple doses or in multiple patient populations.

The importance of protecting the Type I error rate in confirmatory clinical trials is emphasized in several regulatory guidance documents. The European Medicines Agency released the guidance on multiplicity issues in clinical trials (EMA, 2002). More recently, draft guidance documents on this general topic were published by the U.S. Food and Drug Administration (FDA, 2017) and European Medicines Agency (EMA, 2017). These guidelines, especially the FDA guidance, provide a comprehensive discussion of multiplicity issues arising in clinical trials and statistical methods aimed at addressing these issues.

## Overview

---

This chapter deals with a broad class of methods that are used in confirmatory Phase III trials to tackle multiplicity issues. A variety of multiplicity adjustment methods have been proposed in the literature in order to protect the Type I error rate in settings with one or more sources of multiplicity. The chapter describes popular multiple testing procedures (MTPs) and their implementation in SAS. Sections 5.2 and 5.3 discuss basic single-step MTPs (e.g., the Bonferroni procedure) and their improvement into data-driven stepwise MTPs (e.g., the Holm, Hochberg and Hommel procedures). Section 5.4 covers MTPs that rely on prespecified hypothesis testing ordering, such as the fixed-sequence, fallback and chain procedures. Section 5.5 briefly

reviews parametric methods for multiplicity adjustment, i.e., the Dunnett procedure and its stepwise versions. Finally, Section 5.6 introduces a class of gatekeeping procedures that have been developed to perform multiplicity adjustments in clinical trials with two or more sources of multiplicity, e.g., in trials with several hierarchically ordered endpoints and multiple dose-placebo comparisons.

Several recent review papers and book chapters provide a survey of multiplicity adjustments commonly used in clinical trial applications. See Hochberg and Tamhane (1987), Hsu (1996), Dmitrienko et al. (2009), Bretz et al. (2009), Westfall and Bretz (2010), and Dmitrienko and Tamhane (2011) for a more detailed discussion of the general theory of multiple comparisons and different classes of MTPs. Westfall et al. (2011) and Dmitrienko and D'Agostino (2013) provide an overview of MTPs with examples of SAS implementation.

The SAS code and data sets included in this chapter are available on the book's website at <http://support.sas.com/publishing/authors/dmitrienko.html>.

## **Weak and strong control of familywise error rate**

The concept of a Type I error rate originates in the problem of testing a single hypothesis. It is defined as the probability of rejecting the hypothesis when it is true. When multiple hypotheses are being tested, the concept of a Type I error rate can be generalized as the *familywise error rate* (FWER), i.e., the probability of incorrectly rejecting at least one true null hypothesis in the family of hypotheses being tested. There are two definitions of the FWER control: *weak* and *strong* control. It is important to understand what inferences are intended to be performed in order to choose the right definition of the FWER control and an appropriate MTP. To illustrate, consider the following example that will be used throughout Sections 5.2, 5.3 and 5.4.

**EXAMPLE: Case study 1 (Dose-finding hypertension trial)**

Suppose that a dose-finding trial has been conducted to compare low, medium, and high doses of a new antihypertensive drug (labeled L, M, and H) to placebo (labeled P). The primary efficacy variable is diastolic blood pressure. Doses that provide a clinically relevant mean reduction in diastolic blood pressure will be declared efficacious.

Let  $\mu_P$  denote the mean reduction in diastolic blood pressure in the placebo group. Similarly,  $\mu_L$ ,  $\mu_M$ , and  $\mu_H$  denote the mean reduction in diastolic blood pressure in the low, medium, and high dose groups, respectively. The following null hypotheses are tested in the trial:

$$H_L = \{\mu_P - \mu_L \leq \delta\}, \quad H_M = \{\mu_P - \mu_M \leq \delta\}, \quad H_H = \{\mu_P - \mu_H \leq \delta\}.$$

where  $\delta$  represents a clinically significant improvement over placebo. In the context of superiority testing, the clinically significant improvement is set to 0.

Considering the superiority testing framework, the null hypothesis of equality of  $\mu_P$ ,  $\mu_L$ ,  $\mu_M$ , and  $\mu_H$  (known as *the global null hypothesis*) can be tested in the hypertension trial example using the usual *F*-test. The *F*-test is known to preserve the FWER in the *weak sense*, which means that it controls the likelihood of rejecting the global null hypothesis when all individual hypotheses are simultaneously true.

The weak control of the FWER is appropriate only if we want to make a statement about the global null hypothesis. In order to test the effects of the individual doses on diastolic blood pressure, we need to control the probability of erroneously rejecting any true null hypothesis regardless of which and how many null hypotheses are true.

This is referred to as the *strong control* of the FWER. For example, the procedure proposed by Dunnett (1955) can be used to test the global null hypothesis and, at the same time, provides information about the individual dose-placebo comparisons, i.e., it can be used to test the null hypotheses  $H_L$ ,  $H_M$ , and  $H_H$ . The Dunnett procedure controls the FWER in the strong sense at a predefined  $\alpha$  level (e.g., at a two-sided  $\alpha = 0.05$ ). Suppose, for example, that  $H_L$  and  $H_M$  are both true (which means that  $\mu_P \leq \mu_L$  and  $\mu_P \leq \mu_M$ ), whereas  $H_H$  is false (and thus  $\mu_P > \mu_H$ ). In this case, strong control of the FWER implies that the probability of erroneously rejecting  $H_L$  or  $H_M$  by the Dunnett procedure will not exceed  $\alpha$ .

The importance of strong control of the FWER in confirmatory clinical trials is emphasized in the FDA draft guidance on multiple endpoints in clinical trials (FDA, 2017), and this chapter will focus only on methods that preserve the FWER in the strong sense.

It is worth noting that other definitions of the likelihood of an incorrect decision have been proposed in the literature. This includes the false discovery proportion, false discovery rate, and generalized familywise error rate. For more information on these definitions, see Dmitrienko et al. (2009). Multiple testing procedures that control the false discovery rate or generalized familywise error rate are more powerful and, at the same time, more liberal than those designed to protect the FWER in the strong sense. These procedures are only useful in multiplicity problems with a large number of null hypotheses, e.g., large-scale genetic studies.

## Multiple testing procedures

Several classes of MTPs have been developed over the past 20 to 30 years that have found numerous applications in clinical trials. Selection of the most appropriate approaches to handle multiplicity in a particular setting is typically driven by several factors such as available *clinical information* (i.e., information on relevant logical restrictions or dependencies among the null hypotheses in a multiplicity problem) and *statistical information* (i.e., information on the joint distribution of the hypothesis test statistics). To facilitate the process of reviewing available multiplicity adjustment options, Table 5.1 defines a simple classification scheme with nine categories based on the relevant *logical restrictions* and available *distributional information* in the multiplicity problem arising in a clinical trials.

**TABLE 5.1** Classification of popular multiple testing procedures

Logical restrictions		
Single-step	Data-driven hypothesis ordering	Pre-specified hypothesis ordering
Nonparametric procedures		
Bonferroni	Holm	Fixed-sequence Fallback Bonferroni-based chain
Semiparametric procedures		
Simes Šidák	Hochberg Hommel	NA
Parametric procedures		
Dunnett	Step-down Dunnett Step-up Dunnett	Parametric fallback Parametric chain

### Logical restrictions

To provide more information on how clinical information is used in multiplicity problems, it is shown in Table 5.1 that three important classes of MTPs can be defined based on logical restrictions among the null hypotheses of interest: single-step procedures, stepwise procedures that rely on data-driven hypothesis ordering, and stepwise procedures that are based on predefined hypothesis ordering. Single-step procedures, such as the Bonferroni procedure, test each null hypothesis independently of the other hypotheses. This means that the order in which the null hypotheses are examined is not important, and, in other words, the multiple inferences can be thought of as being performed in a single step. By contrast, stepwise procedures, such as the Holm procedure, test one hypothesis at a time in a sequential manner. As a result, some of the null hypotheses might not be tested at all, i.e., they might be either accepted or rejected by implication. Stepwise procedures are superior to simple single-step MTPs in the sense that they increase the number of rejected null hypotheses without inflating the FWER.

The two classes of stepwise procedures defined above, i.e., procedures that rely on data-driven or pre-specified hypothesis ordering, use relevant clinical information to arrange the null hypotheses of interest in the order of importance. Considering the dose-finding hypertension trial example from Case study 1, it might not be possible for the trial's sponsor to determine at the design stage which dose will be the most effective. Thus, it can be challenging to arrange the null hypotheses of no effect in a logical way. In this situation, it is recommended to consider a stepwise procedure with a data-driven testing algorithm, i.e., the hypotheses are tested in the order determined by the significance of the hypothesis test statistics. The Hochberg procedure serves as an example of a stepwise MTP with a data-driven testing sequence. On the other hand, if reliable clinical information is available to define logical restrictions among the null hypotheses, it is sensible to consider procedures that assume a prespecified testing sequence such as the fixed-sequence procedure.

Single-step procedures are discussed in Section 5.2, and commonly used procedures with a data-driven hypothesis ordering are described in Section 5.3. Section 5.4 reviews procedures for which the order of the hypothesis testing is prespecified.

### Distributional information

The other important factor that influences the choice of MTPs in multiplicity problems is the information on the joint distribution of the test statistics associated with the hypotheses of interest. Depending on the amount of available statistical information, three classes of MTPs can be defined: two classes of *p*-value-based procedures (*nonparametric* and *semiparametric* procedures) as well as a class of *parametric* procedures. Testing algorithms used in nonparametric procedures are set up based on univariate *p*-values computed from the hypothesis test statistics without imposing any distributional assumptions. By contrast, semiparametric procedures make some distributional assumptions. However, a full specification of the joint distribution of the test statistics is not required. Parametric procedures require a fully specified joint distribution of the hypothesis test statistics, i.e., the statistics might be assumed to follow a multivariate normal or multivariate *t* distribution. Examples of nonparametric and semiparametric procedures are given in Sections 5.2, 5.3 and 5.4. Parametric procedures are discussed in Section 5.5.

Of note, other classes of MTPs have been proposed, including resampling-based procedures. These procedures do not make assumptions about the joint distribution of the test statistics. Instead, they estimate the joint distribution using bootstrap or permutation methods. Resampling-based procedures only achieve strong FWER control when the sample size approaches infinity. Therefore, they are not used in

confirmatory clinical trials (FDA, 2017). For this reason, MTPs from this class will not be discussed in this chapter. For more information on resampling-based procedures, see Westfall and Young (1993) and Dmitrienko et al. (2009).

## 5.2 Single-step procedures

---

As noted above, there is an important distinction between *single-step* MTPs that test each null hypothesis independently of the other hypotheses and *stepwise* MTPs that rely on sequential testing algorithms with data-driven or predefined hypothesis ordering. This section focuses on the basic single-step procedures, namely, the Bonferroni, Šidák, and Simes adjustments. It is important to note that these procedures might not be of much interest by themselves due to lack of power or other limitations. But, as shown in Sections 5.3 and 5.4, they provide a foundation for defining more efficient procedures.

The following notation will be used in this and subsequent sections. Suppose we plan to carry out  $m$  significance tests corresponding to a family of null hypotheses of no effect denoted by  $H_1, \dots, H_m$ . The global null hypothesis is defined as the intersection of  $H_1, \dots, H_m$ , i.e.,

$$H_1 \cap H_2 \cap \dots \cap H_m.$$

Let  $p_1, \dots, p_m$  denote the individual  $p$ -values generated by the significance tests. The ordered  $p$ -values  $p_{(1)}, \dots, p_{(m)}$  are defined in such a way that

$$p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}.$$

We wish to define a simultaneous test for the family of null hypotheses. The test will be based on a suitable adjustment for multiplicity to keep the FWER at the prespecified  $\alpha$  level, e.g.,  $\alpha = 0.05$  (two-sided). Multiplicity adjustments are performed by modifying the individual decision rules, i.e., by adjusting either the individual  $p$ -values or significance levels ( $p$ -values are adjusted upward or significance levels are adjusted downward). To define adjusted  $p$ -values, we adopt the definition proposed by Westfall and Young (1993). According to this definition, the adjusted  $p$ -value is equal to the smallest significance level for which we would reject the corresponding null hypothesis.

### 5.2.1 Bonferroni and Šidák methods

The Bonferroni and Šidák procedures serve as examples of basic nonparametric and semiparametric multiplicity adjustment methods, respectively. These methods are available in a large number of SAS procedures (for example, PROC GLM, PROC MIXED, and PROC MULTTEST), and use very simple rules. The Bonferroni procedure rejects  $H_i$  if  $p_i \leq \alpha/m$ , and the Šidák procedure rejects  $H_i$  if  $p_i \leq 1 - (1 - \alpha)^{1/m}$ , where  $i = 1, \dots, m$ . The individual adjusted  $p$ -values for the two procedures are given by

$$\tilde{p}_i = mp_i \text{ (Bonferroni)}, \quad \tilde{p}_i = 1 - (1 - p_i)^m \text{ (Šidák)}, \quad i = 1, \dots, m.$$

Using the Bonferroni inequality, it is easy to show that the Bonferroni procedure is a nonparametric procedure that controls the FWER in the strong sense for any joint distribution of the raw  $p$ -values. The Šidák procedure is a semiparametric procedure

that preserves the FWER under additional distributional assumptions. Šidák (1967) demonstrated that the FWER is preserved if the hypothesis test statistics are either independent or follow a multivariate normal distribution. Holland and Copenhaver (1987) described a broad set of assumptions under which the Šidák procedure controls the FWER, for example, when the test statistics follow a multivariate  $t$  distribution and some other distributions.

The Bonferroni and Šidák procedures can also be used to test the global null hypothesis of no effect. The global hypothesis is rejected whenever any of the individual null hypotheses is rejected. This means that the Bonferroni global test rejects the global hypothesis if  $p_i \leq \alpha/m$  for at least one  $i = 1, \dots, m$ . Likewise, the Šidák global test rejects the global hypothesis if  $p_i \leq 1 - (1 - \alpha)^{1/m}$  for at least one  $i = 1, \dots, m$ .

The adjusted  $p$ -values associated with the Bonferroni and Šidák global tests are given by

$$\begin{aligned}\tilde{p}_B &= m \min(p_1, \dots, p_m) \text{ (Bonferroni),} \\ \tilde{p}_S &= 1 - (1 - \min(p_1, \dots, p_m))^m \text{ (Šidák).}\end{aligned}$$

In other words, the Bonferroni and Šidák global tests reject the global null hypothesis if  $\tilde{p}_B \leq \alpha$  and  $\tilde{p}_S \leq \alpha$ , respectively. Although it might not be immediately obvious, the Bonferroni and Šidák global tests are more important than the corresponding MTPs. It will be shown in Section 5.3 that, using these global tests, we can construct closed testing procedures that are uniformly more powerful than the Bonferroni and Šidák procedures.

**It is easy to show that the Šidák correction is uniformly better than the Bonferroni correction (Hsu, 1996, Section 1.3.5).** The difference between these two corrections is rather small when the raw  $p$ -values are small. The two procedures are known to be very conservative when the individual test statistics are highly correlated. **The adjusted  $p$ -values generated by the Bonferroni and Šidák procedures are considerably larger than they need to be to maintain the FWER at the desired level.**

To provide an illustration, consider Case study 1. Recall that  $\mu_P$  denotes the mean reduction in diastolic blood pressure in the placebo group and  $\mu_L$ ,  $\mu_M$  and  $\mu_H$  denote the mean reduction in diastolic blood pressure in the low, medium, and high dose groups, respectively. Negative values of  $\mu_L$ ,  $\mu_M$ , and  $\mu_H$  indicate an improvement in diastolic blood pressure. The three null hypotheses of no treatment difference between each dose and placebo to be tested in the trial were defined in Section 5.1 (the hypotheses are denoted by  $H_L$ ,  $H_M$ , and  $H_H$ ). **The treatment means are compared using the two-sample  $t$ -test.** Table 5.2 shows two-sided  $p$ -values generated by the three dose-placebo tests under four scenarios.

**TABLE 5.2 Dose-placebo comparisons in Case study 1**

Comparison	L vs P	M vs P	H vs P
Scenario 1	$p_L = 0.047$	$p_M = 0.0167$	$p_H = 0.015$
Scenario 2	$p_L = 0.047$	$p_M = 0.027$	$p_H = 0.015$
Scenario 3	$p_L = 0.053$	$p_M = 0.026$	$p_H = 0.017$
Scenario 4	$p_L = 0.022$	$p_M = 0.026$	$p_H = 0.017$

Program 5.1 computes adjusted  $p$ -values produced by the Bonferroni and Šidák procedures under Scenario 1. The two procedures are requested by the BONFERRONI and SIDAK options in PROC MULTTEST.

### PROGRAM 5.1 Bonferroni and Šidák procedures in Case study 1

```

data antihyp1;
    input test $ raw_p @@;
    datalines;
        L 0.047 M 0.0167 H 0.015
    run;

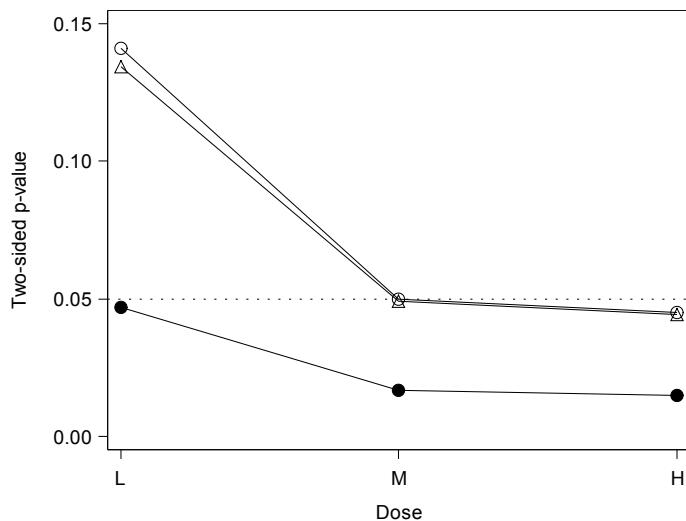
proc multtest pdata=antihyp1 bonferroni sidak out=adjp;
run;

proc sgplot data=adjp noautolegend;
    series x=test y=raw_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=raw_p /
        markerattrs=(symbol=circlefilled color=black size=2pct);
    series x=test y=bon_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=bon_p /
        markerattrs=(symbol=circle color=black size=2pct);
    series x=test y=sid_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=sid_p /
        markerattrs=(symbol=triangle color=black size=2pct);
    refline 0.05 / lineattrs=(pattern=dot color=black thickness=1);
    yaxis label="Two-sided p-value" values=(0 to 0.15 by 0.05)
        labelattrs=(family="Arial" size=9pt);
    xaxis label="Dose" labelattrs=(family="Arial" size=9pt);
run;

```

The result of Program 5.1 is shown in Figure 5.1. The figure displays the adjusted *p*-values produced by the Bonferroni and Šidák procedures plotted along with the

**Figure 5.1**  
Case study 1



*Bonferroni and Šidák procedures. Raw p-value (dot), Bonferroni-adjusted p-value (circle), and Šidák-adjusted p-value (triangle).*

corresponding raw  $p$ -values. The figure indicates that the raw  $p$ -values associated with all three dose-placebo comparisons are significant at the two-sided 5% level, yet only the high dose is significantly different from placebo after the Bonferroni adjustment for multiplicity ( $p = 0.045$ ). The Šidák-adjusted  $p$ -values are consistently less than the Bonferroni-adjusted  $p$ -values. However, the difference is very small. The Šidák-adjusted  $p$ -value for the medium dose versus placebo comparison is marginally significant ( $p = 0.0493$ ), whereas the corresponding Bonferroni-adjusted  $p$ -value is only a notch greater than 0.05 ( $p = 0.0501$ ).

Despite the conservative nature of the Bonferroni adjustment, it has been shown in the literature that the Bonferroni method cannot be improved (Hommel, 1983). We can find fairly exotic  $p$ -value distributions for which the Bonferroni inequality turns into an equality. Therefore, it is impossible to construct a single-step MTP that will be uniformly more powerful than the simple Bonferroni procedure. The only way to improve the Bonferroni method is by making additional assumptions about the joint distribution of the individual  $p$ -values. Examples of MTPs that rely on certain distributional assumptions (e.g., semiparametric and fully parametric procedures) will be given later in this subsection.

### 5.2.2 Simes method

Unlike the Bonferroni and Šidák methods, the Simes method can be used only for testing the global null hypothesis. We will show in Section 5.3 that the Simes method can be extended to perform inferences on individual null hypotheses and demonstrate how to carry out the extended procedures using PROC MULTTEST and custom macros.

The Simes method is closely related to that proposed by Rüger (1978). Rüger noted that the Bonferroni global test can be written as

$$\text{Reject the global null hypothesis } H_1 \cap H_2 \cap \dots \cap H_m \text{ if } p_{(1)} \leq \alpha/m.$$

and uses only a limited amount of information from the sample. It relies on the most significant  $p$ -value and ignores the rest of the  $p$ -values. Rüger described a family of generalized Bonferroni global tests based on the ordered  $p$ -values  $p_{(1)}, \dots, p_{(m)}$ . He showed that the global null hypothesis can be tested using any of the following tests:

$$p_{(1)} \leq \alpha/m, \quad p_{(2)} \leq 2\alpha/m, \quad p_{(3)} \leq 3\alpha/m, \quad \dots, \quad p_{(m)} \leq \alpha.$$

For example, we can test the global null hypothesis by comparing  $p_{(2)}$  to  $2\alpha/m$  or  $p_{(3)}$  to  $3\alpha/m$  as long as we prespecify which one of these tests will be used. Any one of Rüger's global tests controls the FWER in the strong sense for arbitrary dependence structures. However, each test suffers from the same problem as the Bonferroni global test: Each individual test is based on only one  $p$ -value.

Simes (1986) developed a procedure for combining the information from  $m$  individual tests. The Simes global test rejects the global null hypothesis if

$$p_{(i)} \leq i\alpha/m \text{ for at least one } i = 1, \dots, m.$$

Simes demonstrated that his global test is exact in the sense that its size equals  $\alpha$  if  $p_1, \dots, p_m$  are independent. The adjusted  $p$ -value for the global hypothesis associated with this test is equal to

$$\tilde{p}_{SIM} = m \min(p_{(1)}/1, p_{(2)}/2, \dots, p_{(m)}/m).$$

Since  $\tilde{p}_B = mp_{(1)}$ , it is clear that  $\tilde{p}_{SIM} \leq \tilde{p}_B$ . Thus, the Simes global test is uniformly more powerful than the Bonferroni global test.

The Simes test achieves higher power by assuming that the individual  $p$ -values are independent. Naturally, this assumption is rarely satisfied in clinical applications. What happens if the  $p$ -values are correlated? Sarkar and Chang (1997) and Sarkar (1998) initially established strong control of the FWER for families of multivariate distributions characterized by properties of positive dependence and multivariate total positivity of order two. More recently, Sarkar (2008) demonstrated that the Simes global test preserves the FWER if the test statistics follow any multivariate normal distribution with non-negative pairwise correlation coefficients. This setting includes, for example, the dose-finding hypertension trial example from Case study 1 where the test statistics are positively correlated because they are derived from comparisons of three treatment groups to a common control group. If the Simes global test is applied to multiplicity problems with negatively correlated test statistics, the FWER might be inflated.

### 5.2.3 Summary

This section described simple multiplicity adjustment strategies known as single-step methods. Single-step procedures test each null hypothesis of interest independently of the other hypotheses. Therefore, the order in which the null hypotheses are examined becomes unimportant. Single-step MTPs are very easy to implement and have enjoyed much popularity in clinical applications.

The following three single-step MTPs were discussed in the section:

- The Bonferroni procedure controls the FWER for any joint distribution of the marginal  $p$ -values but is known to be rather conservative. Despite the conservative nature of the Bonferroni method, no single-step MTP is uniformly more powerful than the Bonferroni procedure. More powerful MTPs can be constructed only if we are willing to make additional assumptions about the joint distribution of  $p$ -values associated with the null hypotheses of interest.
- The Šidák procedure is uniformly more powerful than the Bonferroni procedure. However, its size depends on the joint distribution of the marginal  $p$ -values and can exceed the nominal level. The Šidák procedure controls the FWER when the test statistics are independent or follow a multivariate normal distribution. The Šidák procedure provides a fairly small improvement over the Bonferroni procedure in Case study 1 as well as general multiplicity problems. The Šidák procedure and related MTPs will not be further discussed in this chapter.
- The Simes test can be used only for testing the global null hypothesis. The Simes global test is more powerful than the Bonferroni global test but does not always preserve the FWER. Its size is known to be no greater than the nominal level when the individual test statistics follow any multivariate normal distribution with non-negative correlation coefficients. As will be shown in Section 5.3, the Simes global test plays a central role in defining powerful Simes-based stepwise procedures for addressing multiplicity in confirmatory clinical trials.

Multiplicity adjustments based on the Bonferroni and Šidák procedures are implemented in PROC MULTTEST and also available in other SAS procedures, e.g., PROC GLM and PROC MIXED.

## 5.3 Procedures with a data-driven hypothesis ordering

---

This section will introduce MTPs that test the null hypotheses in a stepwise manner, as opposed to the single-step approach introduced in Section 5.2. The testing sequence is not *a priori*, specified and the hypotheses are tested in the order determined by the significance of the hypothesis test statistics or, equivalently, *p*-values.

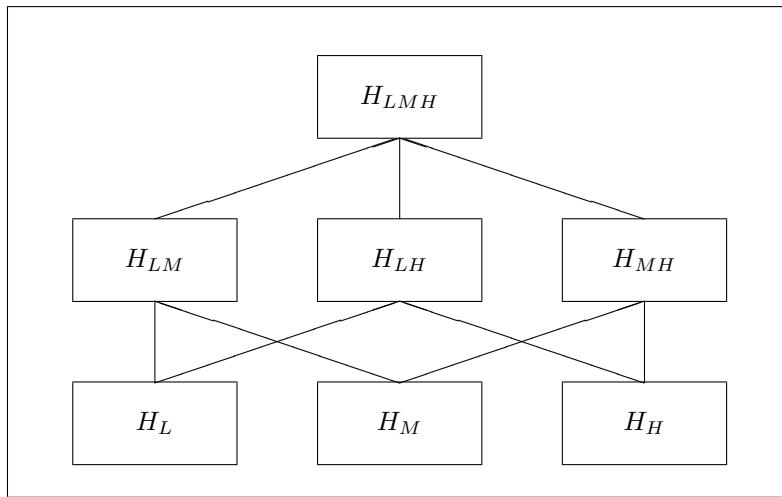
These stepwise procedures are based on a fundamental principle, known as the *closure principle*, which was formulated by Marcus et al. (1976) and has since provided mathematical foundation for numerous multiple testing methods. MTPs constructed using this principle are called closed testing procedures. It would not be an exaggeration to say that virtually all MTPs are either derived using this principle or can be rewritten as closed testing procedures. In fact, Liu (1996) showed that any single-step or stepwise MTP based on marginal *p*-values can be formulated as a closed testing procedure. This means that we can always construct a closed testing procedure that is at least as powerful as any single-step or stepwise MTP. More information on the closure principle is provided below and can also be found in Dmitrienko et al. (2009).

### 5.3.1 Closure principle

The closure principle is based on a hierarchical representation of a multiplicity problem. As an illustration, consider the three null hypotheses  $H_L$ ,  $H_M$ , and  $H_H$  tested in the dose-finding hypertension trial from Case study 1. To apply the closure principle to this multiple testing problem, we need to construct what is known as the *closed family of hypotheses* associated with the three original hypotheses. This is accomplished by forming all possible intersections of  $H_L$ ,  $H_M$ , and  $H_H$ . In a general multiplicity problem involving  $m$  null hypotheses, the closed family is formed of  $2^m - 1$  intersection hypotheses. In our example, the closed family will contain the following  $2^3 - 1 = 7$  intersection hypotheses:

1. Three original hypotheses:  $H_L$ ,  $H_M$  and  $H_H$ .
2. Three intersection hypotheses containing two original hypotheses:  $H_L \cap H_M$ ,  $H_L \cap H_H$  and  $H_M \cap H_H$ .
3. One intersection hypothesis containing three original hypotheses:  $H_L \cap H_M \cap H_H$ .

The next step is to link the intersection hypotheses. The links are referred to as *implication relationships*. A hypothesis that contains another hypothesis is said to imply it. For example,  $H_L \cap H_M \cap H_H$  implies  $H_L \cap H_M$ , which, in turn, implies  $H_L$ . Most commonly, implication relationships are displayed using diagrams similar to that shown in Figure 5.2. Each box in this figure represents a hypothesis in the closed family and is connected to the boxes corresponding to the hypotheses that it contains. Note that  $H_{LMH}$  denotes the intersection hypothesis  $H_L \cap H_M \cap H_H$ ;  $H_{LM}$  denotes the intersection hypothesis  $H_L \cap H_M$ ; and so on. The hypothesis at the top is the intersection of the three original hypotheses. Therefore, it implies  $H_{LM}$ ,  $H_{LH}$ , and  $H_{MH}$ . Likewise, the three boxes at the second level are connected to the boxes representing the original hypotheses  $H_L$ ,  $H_M$ , and  $H_H$ . Alternatively, we can put together a list of all intersection hypotheses in the closed family implying the three original hypotheses (Table 5.3).



**Figure 5.2**  
**Case study 1**

Implication relationships in the closed family of null hypotheses in the hypertension trial from Case study 1. Each box in this diagram represents a hypothesis in the closed family and is connected to the boxes that correspond to the hypotheses it implies.

**TABLE 5.3** Intersection hypotheses implying the original hypotheses

Original hypothesis	Intersection hypotheses implying the original hypothesis
$H_L$	$H_L, H_{LM}, H_{LH}, H_{LMH}$
$H_M$	$H_M, H_{LM}, H_{MH}, H_{LMH}$
$H_H$	$H_H, H_{LH}, H_{MH}, H_{LMH}$

The closure principle states that we can control the FWER in the strong sense by using the following multiplicity adjustment method.

**Closure principle.** Test each hypothesis in the closed family using a suitable  $\alpha$ -level significance test that controls the error rate at the hypothesis level. A hypothesis is rejected if its associated test and all tests associated with hypotheses that imply the hypothesis in question are significant.

According to the closure principle, to reject any of the original null hypotheses in the left column of Table 5.3, we need to test and reject all associated intersection hypotheses shown in the right column. If any of the intersection hypotheses is accepted, then all hypotheses implied by it are accepted without testing.

Any valid significance test can be used to test intersection hypotheses in the closed family as long as its size does not exceed  $\alpha$  at the hypothesis level. We can carry out the global  $F$ -test if the individual test statistics are independent or the Bonferroni global test if they are not. Each of these tests will result in a different closed testing procedure. However, as shown by Marcus et al. (1976), any MTP based on the closure principle controls the FWER in the strong sense at the prespecified  $\alpha$  level as long as an  $\alpha$  level test is used to test each intersection hypothesis.

The closure principle provides statisticians with a very powerful tool for addressing multiplicity problems in numerous settings. The only major disadvantage of closed testing procedures is that it is generally difficult, if not impossible, to construct associated simultaneous confidence intervals for parameters of interest. Other principles have been introduced in the literature for defining MTPs, for example, the partitioning principle (Stefansson et al., 1988; Finner and Strassburger,

2002). The *partitioning principle* can be used to define simultaneous confidence sets associated with popular MTPs (Hsu and Berger, 1999), and the principle has also been applied to more advanced multiplicity problems that arise in clinical trials (Liu and Hsu, 2009). Another important principle is the *sequential rejection principle* (Goeman and Solari, 2010), which gives a different theoretical perspective on many popular MTPs and emphasizes sequential testing algorithms.

### 5.3.2 Decision matrix algorithm

In this subsection, we will introduce the *decision matrix algorithm* for implementing general closed testing procedures. This algorithm streamlines the decision-making process and simplifies the computation of adjusted  $p$ -values associated with the intersection and original hypotheses. See Dmitrienko et al. (2003) for details and examples.

As an illustration, suppose we wish to enhance the Bonferroni correction by using the closure principle. To do so, the Bonferroni global test is applied to each intersection hypothesis to build a nonparametric closed testing procedure known as the Holm procedure (Holm, 1979). It will be shown later that the Holm procedure has a simple stepwise representation and that it is not necessary to test all intersection hypotheses in the closed family. However, such a shortcut is not always available, and the decision matrix algorithm serves as a powerful tool for general implementation of closed testing procedures. The decision matrix algorithm will also be applied to construct closed testing procedures for setting up gatekeeping procedures in Section 5.6.

### Holm procedure

Choose an arbitrary intersection hypothesis  $H$  in the closed family (see Table 5.3). The hypothesis will be tested using the Bonferroni global test, i.e., the following decision rule will be used:

Compute the  $p$ -value associated with the Bonferroni global test. The  $p$ -value is equal to the most significant  $p$ -value corresponding to the original hypotheses implied by  $H$  times the number of original hypotheses implied by  $H$ . Denote the obtained  $p$ -value by  $\tilde{p}_B$ . Reject  $H$  if  $\tilde{p}_B \leq \alpha$ .

For example, consider the intersection hypothesis  $H_{LM}$ . This intersection hypothesis implies two original hypotheses, namely,  $H_L$  and  $H_M$ . Therefore, the Bonferroni-adjusted  $p$ -value is given by  $\tilde{p}_B = 2 \min(p_L, p_M)$ . The hypothesis will be rejected if  $\tilde{p}_B \leq \alpha$ .

By the Bonferroni inequality, the size of the Bonferroni global test is no greater than  $\alpha$ . This means that we have constructed a family of  $\alpha$ -level significance tests for each hypothesis in the closed family. Therefore, applying the closure principle yields an MTP for the original hypotheses  $H_L$ ,  $H_M$ , and  $H_H$  that protect the FWER in the strong sense.

Consider the null hypotheses tested in Case study 1. Table 5.4 presents a decision matrix for the Holm procedure. It summarizes the algorithm for computing the adjusted  $p$ -values that are later used in testing the significance of individual dose-placebo comparisons.

There are 7 rows in Table 5.4, each corresponding to a single intersection hypothesis in the closed family. For each of these intersection hypotheses, the right panel of the table identifies the implied hypotheses, and the central panel displays the

**TABLE 5.4 Decision matrix for the Holm procedure**

Intersection hypothesis	P-value	Implied hypotheses		
		$H_L$	$H_M$	$H_H$
$H_{LMH}$	$p_{LMH} = 3 \min(p_L, p_M, p_H)$	$p_{LMH}$	$p_{LMH}$	$p_{LMH}$
$H_{LM}$	$p_{LM} = 2 \min(p_L, p_M)$	$p_{LM}$	$p_{LM}$	0
$H_{LH}$	$p_{LH} = 2 \min(p_L, p_H)$	$p_{LH}$	0	$p_{LH}$
$H_L$	$p_L = p_L$	$p_L$	0	0
$H_{MH}$	$p_{MH} = 2 \min(p_M, p_H)$	0	$p_{MH}$	$p_{MH}$
$H_M$	$p_M = p_M$	0	$p_M$	0
$H_H$	$p_H = p_H$	0	0	$p_H$

formula for computing the associated  $p$ -values denoted by

$$p_{LMH}, p_{LM}, p_{LH}, p_{MH}, p_L, p_M, p_H.$$

In order to make inferences about the three original hypotheses, we first compute these  $p$ -values and populate the right panel of the table. Westfall and Young (1993) stated that the adjusted  $p$ -value for each original hypothesis is equal to the largest  $p$ -value associated with the intersection hypotheses that imply it. It means that we can obtain the adjusted  $p$ -values for  $H_L$ ,  $H_M$ , and  $H_H$  by computing the largest  $p$ -value in the corresponding column in the right panel. For example, the adjusted  $p$ -value for the low dose versus placebo comparison is equal to

$$\max(p_L, p_{LM}, p_{LH}, p_{LMH}).$$

Program 5.2 computes the Holm-adjusted  $p$ -values using the decision matrix algorithm based on the two-sided raw  $p$ -values that correspond to Scenario 1 in Table 5.2.

### PROGRAM 5.2 Holm procedure (decision matrix approach) in Case study 1

```

proc iml;
  use antihyp1;
  read all var {raw_p} into p;
  h=j(7,3,0);
  decision_matrix=j(7,3,0);
  adjusted_p=j(1,3,0);
  do i=1 to 3;
    do j=0 to 6;
      k=floor(j/2**((3-i));
      if k/2=floor(k/2) then h[j+1,i]=1;
    end;
  end;
  do i=1 to 7;
    decision_matrix[i,]=h[i,]*sum(h[i,])*min(p[loc(h[i,])]);
  end;
  do i=1 to 3;
    adjusted_p[i]=max(decision_matrix[,i]);
  end;
  title={"L vs P", "M vs P", "H vs P"};
  print decision_matrix[colname=title];
  print adjusted_p[colname=title];
quit;

```

**Output from  
Program 5.2**

DECISION_MATRIX		
L vs P	M vs P	H vs P
0.045	0.045	0.045
0.0334	0.0334	0
0.03	0	0.03
0.047	0	0
0	0.03	0.03
0	0.0167	0
0	0	0.015

ADJUSTED_P		
L vs P	M vs P	H vs P
0.047	0.045	0.045

The DECISION\_MATRIX table in Output 5.2 represents the right panel of Table 5.4 and serves as a helpful tool for visualizing the decision-making process behind the closed testing procedure. As was noted earlier, the adjusted  $p$ -values for  $H_L$ ,  $H_M$ , and  $H_H$  can be obtained by computing the maximum over all  $p$ -values in the corresponding column. The computed  $p$ -values are shown at the bottom of Output 5.2.

### 5.3.3 Stepwise procedures

In certain cases, the decision matrix algorithm can be significantly simplified. In those cases, the significance testing can be carried out in a straightforward stepwise manner without examining all implication relationships. Consider a general family of null hypotheses  $H_1, \dots, H_m$ . As noted above, stepwise procedures rely on a data-driven testing sequence, namely, the ordered hypotheses  $H_{(1)}, \dots, H_{(m)}$ , corresponding to the ordered  $p$ -values  $p_{(1)}, \dots, p_{(m)}$ . Two classes of stepwise procedures, *step-down* and *step-up* procedures, are defined as follows:

- Step-down procedures start testing the hypothesis corresponding to the most significant  $p$ -value and continue testing other hypotheses in a sequentially rejective fashion until a certain hypothesis is accepted or all hypotheses are rejected. If a hypothesis is accepted, testing stops, and the remaining hypotheses are accepted by implication. The Holm procedure is an example of a step-down testing procedure.
- Step-up procedures implement the hypothesis testing starting from the opposite direction from the least significant  $p$ -value to the most significant one. As opposed to step-down procedures, once a step-up procedure rejects a hypothesis, it rejects the rest of the hypotheses by implication. In addition, the testing does not stop after a hypothesis is accepted. The Hommel and Hochberg procedures, introduced below, are examples of step-up testing procedures.

#### Holm stepwise procedure: Step-down algorithm

The Holm procedure was introduced earlier using the decision matrix algorithm. A stepwise algorithm is proposed below for the Holm procedure that enables a simpler decision rule. In fact, using the decision matrix algorithm, we need to go through  $2^m - 1$  steps (all intersection hypotheses in the closed family) in order to make final

inference regarding the  $m$  original null hypotheses, but a maximum of  $m$  steps are required using the stepwise approach. The Holm stepwise procedure is a step-down procedure. Thus, it begins with the hypothesis associated with the most significant  $p$ -value, i.e., with  $H_{(1)}$ . The complete testing algorithm is as follows:

- Step 1. Reject  $H_{(1)}$  if  $p_{(1)} \leq \alpha/m$ . If  $H_{(1)}$  is rejected, proceed to Step 2; otherwise, testing stops and the remaining hypotheses are automatically accepted.
- Steps  $i = 2, \dots, m - 1$ . Reject  $H_{(i)}$  if  $p_{(i)} \leq \alpha/(m - i + 1)$ . If  $H_{(i)}$  is rejected, proceed to Step  $i + 1$ ; otherwise, testing stops and the remaining hypotheses are automatically accepted.
- Step  $m$ . Reject  $H_{(m)}$  if  $p_{(m)} \leq \alpha$ .

To compare the Holm and Bonferroni testing procedures, note that the Bonferroni procedure tests  $H_1, \dots, H_m$  at the same  $\alpha/m$  level. In contrast, the Holm procedure tests  $H_{(1)}$  at the  $\alpha/m$  level, and the other hypotheses are tested at successively higher significance levels. As a result, the Holm stepwise procedure rejects at least as many (and possibly more) hypotheses as the Bonferroni procedure while maintaining the FWER at the same level.

To illustrate this stepwise testing algorithm, consider Scenario 1 in Table 5.2. Program 5.3 shows how to compute adjusted  $p$ -values produced by the Bonferroni and Holm procedures using PROC MULTTEST. The Bonferroni and Holm procedures are requested by the BONFERRONI and HOLM options.

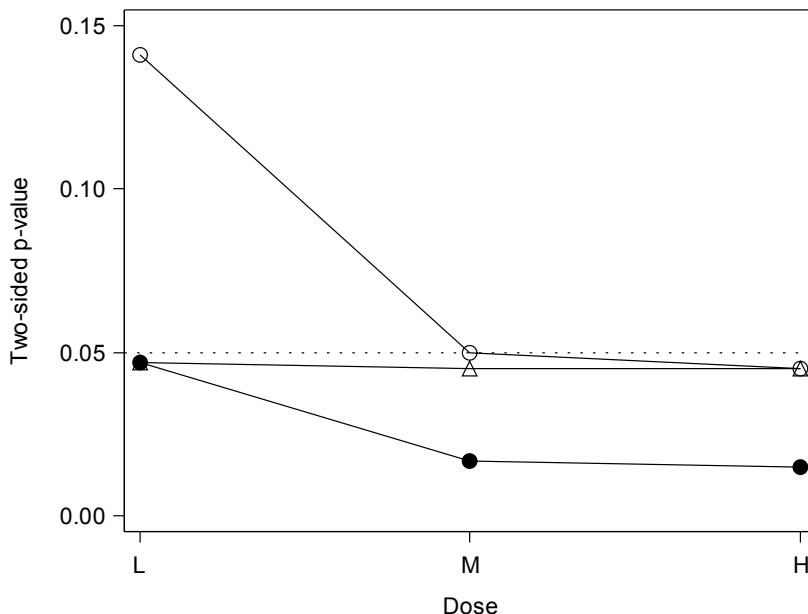
### PROGRAM 5.3    Bonferroni and Holm procedures in Case study 1

```
proc multtest pdata=antihyp1 bonferroni holm out=adjp;
run;

proc sgplot data=adjp noautolegend;
  series x=test y=raw_p /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=test y=raw_p /
    markerattrs=(symbol=circlefilled color=black size=2pct);
  series x=test y=bon_p /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=test y=bon_p /
    markerattrs=(symbol=circle color=black size=2pct);
  series x=test y=stpbon_p /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=test y=stpbon_p /
    markerattrs=(symbol=triangle color=black size=2pct);
  reffline 0.05 / lineattrs=(pattern=dot color=black thickness=1);
  yaxis label="Two-sided p-value" values=(0 to 0.15 by 0.05)
    labelattrs=(family="Arial" size=9pt);
  xaxis label="Dose" labelattrs=(family="Arial" size=9pt);
run;
```

The result of Program 5.3 is displayed in Figure 5.3. Figure 5.3 shows that all three Holm-adjusted  $p$ -values are less than 0.05. (Note that the adjusted  $p$ -value for the low dose versus placebo comparison is equal to the corresponding raw  $p$ -value.) This means that the Holm stepwise procedure has rejected all three null hypotheses, indicating that all doses of the experimental treatment provide a significantly greater reduction in diastolic blood pressure compared to placebo. In contrast, only the high dose is significantly different from placebo according to the Bonferroni correction. Note that the adjusted  $p$ -values produced by PROC MULTTEST are identical to those obtained with the decision matrix algorithm implemented in Program 5.2.

**Figure 5.3**  
Case study 1



*Bonferroni and Holm procedures. Raw p-value (dot), Bonferroni-adjusted p-value (circle), and Holm-adjusted p-value (triangle).*

It is easy to see that a sequential approach can be derived from the Holm stepwise algorithm to simply calculate these adjusted  $p$ -values in  $m$  steps. The adjusted  $p$ -value associated with  $H_{(i)}$ , denoted by  $\tilde{p}_{(i)}$ , is found as follows:

$$\begin{aligned}\tilde{p}_{(1)} &= mp_{(1)}, \\ \tilde{p}_{(i)} &= \max(\tilde{p}_{(i-1)}, (m - i + 1)p_{(i)}), \quad i = 2, \dots, m.\end{aligned}$$

### Hommel procedure: Step-up algorithm

It was shown in the previous subsection that we can improve the performance of the Bonferroni correction by constructing a closed testing procedure based on the Bonferroni global test (i.e., Holm procedure). The same idea can be used to enhance any single-step test for the global null hypothesis described in Section 5.2.

In particular, a closed testing procedure based on the Simes global test was proposed by Hommel (1986, 1988). Since the Simes global test is more powerful than the Bonferroni global test, the Hommel procedure rejects all hypotheses rejected by the Holm procedure and possibly more. However, the improvement comes with a price because the Hommel procedure is a semiparametric procedure and does not always guarantee FWER control. The Hommel procedure protects the FWER only when the Simes global test does, e.g., when the individual test statistics follow a multivariate normal distribution with non-negative correlation coefficients (Sarkar, 2008).

The Hommel procedure can be implemented using the decision matrix algorithm by applying the Simes global test to each intersection hypothesis. However, there is also a stepwise sequential formulation of the Hommel procedure that streamlines the decision making process (Brechenmacher et al., 2011). The Hommel stepwise

algorithm is a step-up procedure, and, as such, it starts by testing the hypotheses corresponding to the least significant  $p$ -value, i.e.,  $H_{(m)}$ , in the following manner:

- Step 1. Accept  $H_{(m)}$  if  $p_{(m)} > \alpha$ . If  $H_{(m)}$  is accepted, proceed to Step 2. Otherwise, testing stops and all hypotheses are rejected.
- Steps  $i = 2, \dots, m - 1$ . Accept  $H_{(m-i+1)}$  if  $p_{(m-i+j)} > j\alpha/i$  for all  $j = 1, \dots, i$ . If  $H_{(m-i+1)}$  is accepted, proceed to Step  $i + 1$ . Otherwise, reject  $H_{(j)}$  if  $p_{(j)} \leq \alpha/(i-1)$ ,  $j = 1, \dots, m - i + 1$  and stop testing.
- Step  $m$ . Accept  $H_{(1)}$  if  $p_{(j)} > j\alpha/m$  for all  $j = 1, \dots, m$  or  $p_{(1)} > \alpha/(m-1)$ . Otherwise, reject  $H_{(1)}$ .

The Hommel procedure is available in PROC MULTTEST and can be requested using the HOMMEL option. Program 5.4 computes the Holm- and Hommel-adjusted  $p$ -values for the individual dose-placebo comparisons in Case study 1 under Scenario 2 shown in Table 5.2.

#### **PROGRAM 5.4    Holm and Hommel procedures in Case study 1**

```

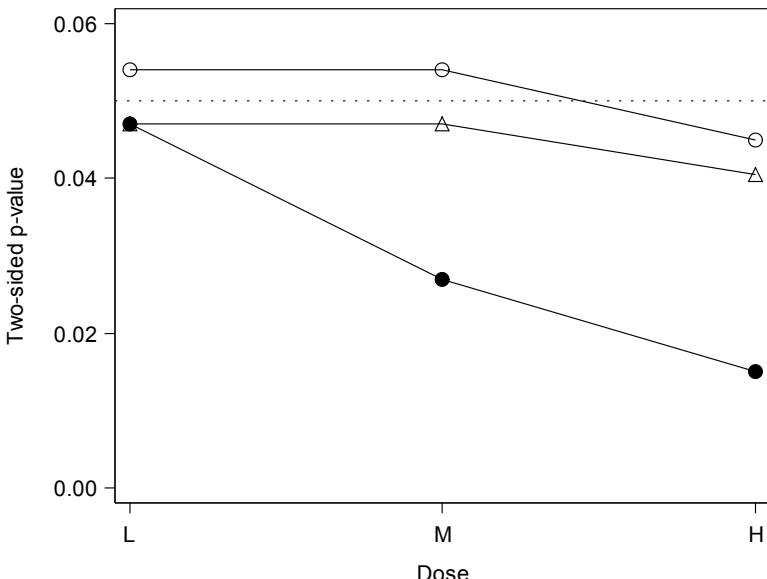
data antihyp2;
    input test $ raw_p @@;
    datalines;
        L 0.047 M 0.027 H 0.015
run;

proc multtest pdata=antihyp2 holm hommel out=adjp;
run;

proc sgplot data=adjp noautolegend;
    series x=test y=raw_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=raw_p /
        markerattrs=(symbol=circlefilled color=black size=2pct);
    series x=test y=stpbond_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=stpbond_p /
        markerattrs=(symbol=circle color=black size=2pct);
    series x=test y=hom_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=hom_p /
        markerattrs=(symbol=triangle color=black size=2pct);
    refline 0.05 / lineattrs=(pattern=dot color=black thickness=1);
    yaxis label="Two-sided p-value" values=(0 to 0.06 by 0.02)
        labelattrs=(family="Arial" size=9pt);
    xaxis label="Dose" labelattrs=(family="Arial" size=9pt);
run;

```

The result of Program 5.4 is displayed in Figure 5.4. Figure 5.4 demonstrates that the Hommel testing procedure has rejected all three null hypotheses. Thus, it is clearly more powerful than the Holm procedure that rejected only one hypothesis. (Note that the Hommel-adjusted  $p$ -value for the low dose versus placebo comparison is equal to the corresponding raw  $p$ -value.) The figure also illustrates an interesting property of the Hommel procedure. Since the Hommel procedure is based on the Simes global test, it rejects all null hypotheses whenever all raw  $p$ -values are significant.



**Figure 5.4**  
Case study 1

Holm and Hommel procedures. Raw p-value (dot), Holm-adjusted p-value (circle), and Hommel-adjusted p-value (triangle).

### Hochberg procedure: Step-up algorithm

Another popular stepwise procedure supported by PROC MULTTEST is the Hochberg procedure (Hochberg, 1988). This MTP is virtually identical to the Holm procedure except for the order in which the null hypotheses are tested. Indeed, the Hochberg procedure examines the ordered  $p$ -values  $p_{(1)}, \dots, p_{(m)}$  starting with the largest one and thus falls in the class of step-up procedures. The detailed testing algorithm is as follows:

- Step 1. Accept  $H_{(m)}$  if  $p_{(m)} > \alpha$ . If  $H_{(m)}$  is accepted, proceed to Step 2; otherwise, testing stops and all hypotheses are rejected.
- Steps  $i = 2, \dots, m-1$ . Accept  $H_{(m-i+1)}$  if  $p_{(m-i+1)} > \alpha/i$ . If  $H_{(m-i+1)}$  is accepted, proceed to Step  $i + 1$ ; otherwise, testing stops and the remaining hypotheses are rejected.
- Step  $m$ . Reject  $H_{(1)}$  if  $p_{(1)} \leq \alpha/m$ .

The Hochberg procedure rejects all hypotheses rejected by the Holm procedure and possibly more, but it is uniformly less powerful than the Hommel procedure (Hommel, 1989). In fact, the Hochberg closed testing procedure is constructed using a simplified conservative version of the Simes global test. This means that in all situations when both of these MTPs can be carried out, the Hommel procedure will offer a power advantage over the Hochberg procedure. Note that the Hommel and Hochberg MTPs are equivalent when only two null hypotheses are tested. The Hochberg procedure is also semiparametric and protects the FWER under the same conditions as the Simes global test. The FWER can potentially exceed  $\alpha$ , but it is no greater than  $\alpha$  when the test statistics follow any multivariate normal distribution with non-negative correlation coefficients.

The Hochberg procedure is available in PROC MULTTEST and can be requested using the HOCHBERG option. Program 5.5 computes the Hochberg- and Hommel-adjusted  $p$ -values for the three null hypotheses tested in Case study 1 under Scenario 3 shown in Table 5.2.

### PROGRAM 5.5 Hochberg and Hommel procedures in Case study 1

```

data antihyp3;
    input test $ raw_p @@;
    datalines;
    L 0.053 M 0.026 H 0.017
run;

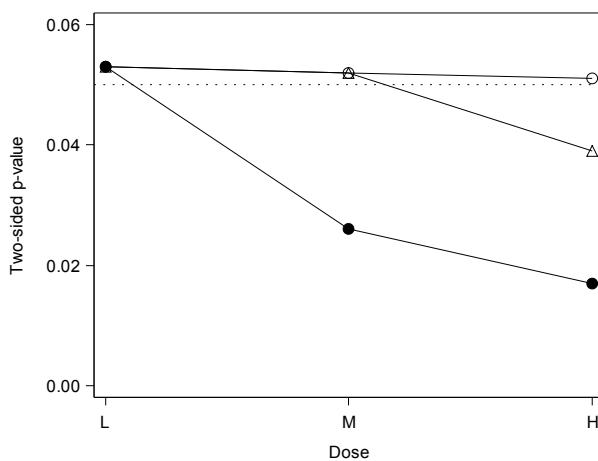
proc multtest pdata=antihyp3 hochberg hommel out=adjp;
run;

proc sgplot data=adjp noautolegend;
    series x=test y=raw_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=raw_p /
        markerattrs=(symbol=circlefilled color=black size=2pct);
    series x=test y=hoc_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=hoc_p /
        markerattrs=(symbol=circle color=black size=2pct);
    series x=test y=hom_p /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=test y=hom_p /
        markerattrs=(symbol=triangle color=black size=2pct);
    refline 0.05 / lineattrs=(pattern=dot color=black thickness=1);
    yaxis label="Two-sided p-value" values=(0 to 0.06 by 0.02)
        labelattrs=(family="Arial" size=9pt);
    xaxis label="Dose" labelattrs=(family="Arial" size=9pt);
run;

```

The result of Program 5.5 is shown in Figure 5.5. We can see from Figure 5.5 that the Hommel procedure has rejected more null hypotheses than the Hochberg procedure. Note that both the Hochberg- and Hommel-adjusted  $p$ -values for the low dose versus placebo comparison are equal to the corresponding raw  $p$ -value.

**Figure 5.5**  
Case study 1



*Hochberg and Hommel procedures. Raw p-value (dot), Hochberg-adjusted p-value (circle), and Hommel-adjusted p-value (triangle).*

### 5.3.4 Weighted procedures

It is important to note that, for all MTPs introduced so far, it has been assumed that hypotheses were equally important, i.e., all hypotheses were equally weighted. In some situations, it can be desirable to assign hypothesis-specific weights to quantify the relative importance of the individual hypotheses in the multiplicity problem. For example, weighted hypotheses can be encountered in dose-finding trials where different weights can be assigned to the dose-placebo comparisons according to the expected effect size at each dose. With this approach, the trial's sponsor may be able to improve the overall power of the MTP.

For example, the Bonferroni procedure introduced in Section 5.2 can be easily extended to a weighted version. Let  $w_1, \dots, w_m$  be the weights corresponding to the hypotheses  $H_1, \dots, H_m$ . Each weight is between 0 and 1, and the weights add up to 1. The individual adjusted  $p$ -values for the weighted Bonferroni procedure are given by

$$\tilde{p}_i = \frac{p_i}{w_i}, \quad i = 1, \dots, m.$$

It is easy to see that, if all weights are equal, i.e.,  $w_i = 1/m$ , the weighted Bonferroni procedure simplifies to the regular Bonferroni procedure.

Similarly, weighted versions of the Holm, Hochberg, and Hommel procedures are available and are defined below using the decision matrix algorithm introduced in Section 5.3.2. Let  $H$  be an intersection hypothesis in the closed family composed of  $n$  null hypotheses, and let  $p_1, \dots, p_n$  and  $w_1, \dots, w_n$  be respectively the raw  $p$ -values and weights corresponding to each null hypothesis within this intersection. Further, the ordered  $p$ -values are denoted by  $p_{(1)} \leq \dots \leq p_{(n)}$ , and their corresponding weights are denoted by  $w_{(1)}, \dots, w_{(n)}$ . The intersection  $p$ -values for the weighted Holm, Hochberg, and Hommel procedures are defined as

$$p(H) = \min_{i=1, \dots, n} \frac{(w_1 + \dots + w_n)p_i}{w_i} \text{ (Holm)}$$

$$p(H) = \min_{i=1, \dots, n} \frac{(w_{(i)} + \dots + w_{(n)})p_{(i)}}{w_{(i)}} \text{ (Hochberg)}$$

$$p(H) = \min_{i=1, \dots, n} \frac{(w_{(1)} + \dots + w_{(n)})p_{(i)}}{w_{(1)} + \dots + w_{(i)}} \text{ (Hommel).}$$

If all weights are equal to each other, the weighted procedures reduce to the regular procedures presented in Section 5.3.3.

Two step-down algorithms have been proposed for implementing the weighted Holm procedure in Holm (1979) and Benjamini and Hochberg (1997). However, stepwise procedures based on the weighted Hochberg procedure cannot be easily derived. For example, as opposed to the regular Hochberg procedure, the weighted Hochberg procedure cannot be formulated as the step-up analogue of the weighted Holm step-down algorithm because the resulting procedure fails to control the FWER. In fact, Tamhane and Liu (2008) highlighted the difficulties for constructing a stepwise version of the weighted Hochberg procedure that protects the FWER. Similarly, the weighted Hommel procedure does not have a stepwise form.

Considering the hypertension trial example from Case study 1, if the high dose is expected to be more efficacious compared to other doses, and the low and medium doses are assumed to be similarly efficacious, the trial's sponsor might choose to put more importance on the high dose and assign the following weights to  $H_L$ ,  $H_M$ , and  $H_H$ :

$$w_L = \frac{1}{4}, \quad w_M = \frac{1}{4}, \quad w_H = \frac{1}{2}.$$

The resulting decision matrix for the weighted Holm procedure is given in Table 5.5. When compared to the decision matrix for the regular Holm procedure in Table 5.4, it can be seen that the multiplicity adjustment strategy is different in that the burden applied to the  $p$ -value corresponding to the comparison of the high dose against placebo,  $p_H$ , is lower at the detriment of the other  $p$ -values.

**TABLE 5.5 Decision matrix for the weighted Holm procedure**

Intersection hypothesis	$P$ -value	Implied hypotheses		
		$H_L$	$H_M$	$H_H$
$H_{LMH}$	$p_{LMH} = \min(4p_L, 4p_M, 2p_H)$	$p_{LMH}$	$p_{LMH}$	$p_{LMH}$
$H_{LM}$	$p_{LM} = 2 \min(p_L, p_M)$	$p_{LM}$	$p_{LM}$	0
$H_{LH}$	$p_{LH} = \min(3p_L, \frac{3}{2}p_H)$	$p_{LH}$	0	$p_{LH}$
$H_L$	$p_L = p_L$	$p_L$	0	0
$H_{MH}$	$p_{MH} = \min(3p_M, \frac{3}{2}p_H)$	0	$p_{MH}$	$p_{MH}$
$H_M$	$p_M = p_M$	0	$p_M$	0
$H_H$	$p_H = p_H$	0	0	$p_H$

It is important to note that PROC MULTTEST does not currently support weighted versions of popular multiple testing procedures, and these procedures can be implemented using the %PvalProc macro. This macro supports the Bonferroni, Holm, Hochberg and Hommel procedures in multiplicity problems with unequally weighted null hypotheses. The arguments of this macro are defined below:

- **in** is the name of the data set with raw  $p$ -values and hypothesis weights. This data set must contain one row per null hypothesis.
- **out** is the name of the data set with the multiplicity-adjusted  $p$ -values for each of the four MTPs.

Program 5.6 uses the %PvalProc macro to compute adjusted  $p$ -values for the weighted Bonferroni, Holm, Hochberg, and Hommel procedures in Case study 1 under Scenario 3. In this example, the weights are chosen as  $w_L = 1/4$ ,  $w_M = 1/4$ , and  $w_H = 1/2$ . The %PvalProc macro requires that the raw  $p$ -values and weights be specified in the **in** data set.

#### PROGRAM 5.6 Weighted procedures in Case study 1

```

data antihyp3;
  input raw_p weight;
  datalines;
  0.053 0.25
  0.026 0.25
  0.017 0.5
  ;
run;

%PvalProc(in=antihyp3,out=adjp);

proc print data=adjp noobs;
  title "Adjusted p-values for weighted procedures";
  var test raw bonferroni holm hochberg hommel;
run;

```

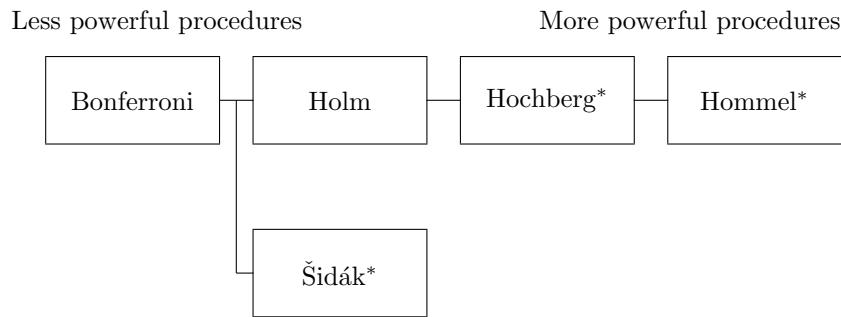
Output from Program 5.6		Adjusted p-values for weighted procedures				
test		RAW	BONFERRONI	HOLM	HOCHBERG	HOMMEL
1		0.053	0.212	0.053	0.053	0.053
2		0.026	0.104	0.052	0.052	0.052
3		0.017	0.034	0.034	0.034	0.034

Output 5.6 displays the adjusted  $p$ -values calculated based on the different weighted procedures. Using all four procedures, only the null hypothesis corresponding to the high dose (Test 3) can be rejected at a two-sided  $\alpha = 0.05$ . Interestingly, none of the hypotheses could have been rejected with any of the regular unweighted procedures, which illustrates the benefit of assigning a greater weight to the high dose in this scenario.

### 5.3.5 Power comparisons

Figure 5.6 displays the relationship among the two single-step and three stepwise MTPs discussed in this and previous sections. Note that this relationship remains the same for the corresponding weighted procedures as long as the procedures use the same set of weights. The multiple procedures shown in Figure 5.6 are arranged in the order of increasing power. The MTPs on the right side are uniformly more powerful than those on the left side. Although the testing procedures derived from the Simes and related MTPs (i.e., Hochberg and Hommel procedures) are more powerful than the Holm and Bonferroni procedures, we need to remember that these procedures do not always control the FWER as explained earlier in this section.

Recall that the Hochberg and Hommel are semiparametric procedures, and, if they are applied to multiplicity problems with negatively correlated test statistics, the FWER might be inflated. However, very limited analytical results are available in the literature to characterize the specific conditions under which the FWER is indeed inflated. Sarkar and Chang (1997) evaluated the FWER associated with the Simes global test in multiplicity problems with three, five, and ten null hypotheses.



**Figure 5.6  
Popular multiple testing procedures**

A comparison of five popular multiple testing procedures. Procedures displayed on the right side are uniformly more powerful than those on the left side. An asterisk indicates that the procedure is semiparametric and controls the FWER under additional assumptions on the joint distribution of the hypothesis test statistics.

The hypothesis test statistics were assumed to follow a multivariate  $t$  distribution, and, with a one-sided  $\alpha = 0.025$ , the highest error rate over a large number of scenarios was 0.0254. This simple example shows that, in the worst-case scenario, the error rate was inflated by less than 2% on a relative scale.

### 5.3.6 Summary

The closure principle introduced in this section provides clinical trial statisticians with a very powerful tool for addressing multiplicity issues and has found numerous applications in clinical trials. This section discussed three popular closed testing procedures that rely on testing algorithms with a data-driven hypothesis ordering:

- The Holm procedure is a nonparametric procedure derived from the Bonferroni global test. Therefore, it controls the FWER for arbitrarily dependent marginal  $p$ -values. The Holm procedure is based on a step-down sequentially rejective algorithm that tests the hypothesis associated with the most significant  $p$ -value at the same level as the Bonferroni procedure. But the other hypotheses are tested at successively higher significance levels. As a consequence, the stepwise Holm procedure is uniformly more powerful than the single-step Bonferroni procedure.
- The Hochberg procedure is a semiparametric procedure with a step-up testing algorithm that examines the least significant  $p$ -value first and then works upward. This procedure is superior to the Holm procedure, but its size can potentially exceed the nominal level. The Hochberg procedure controls the FWER in all situations when the Simes global test does.
- The Hommel procedure is another semiparametric procedure that uses a step-up algorithm. This procedure is uniformly more powerful than both the Holm and Hochberg procedures, and, like the Hochberg procedure, is known to preserve the FWER in the strong sense only when the Simes global test does. For this reason, the Hommel procedure might be favored in all situations when one can carry out the Hochberg procedure.

In addition, it was shown that these procedures have weighted alternatives that enable more flexibility in defining the relative importance of each hypothesis being tested. These procedures can be implemented in PROC MULTTEST if the null hypotheses of interest are equally weighted. The weighted Bonferroni, Holm, Hochberg, and Hommel procedures are supported by the %PvalProc macro.

---

## 5.4 Procedures with a prespecified hypothesis ordering

Section 5.3 introduced several stepwise MTPs that rely on a data-driven hypothesis ordering. This section will define stepwise MTPs that test hypotheses in an order that is prespecified at the trial design stage and does not depend on the observed data.

There is a very important distinction between the two approaches to setting up stepwise procedures. MTPs introduced in Section 5.3 are adaptive in the sense that the order in which the null hypotheses of interest are tested is driven by the data. The hypotheses that are likely to be false are tested first, and testing ceases when none of the untested hypotheses appears to be false. For example, the Holm procedure examines the individual  $p$ -values from the most significant ones to the least significant ones. The tests stops when all remaining  $p$ -values are too large to be significant. Procedures from the second class rely heavily on the assumption that the order in which the null hypotheses are tested is predetermined. As a result, we run a risk of accepting false hypotheses because they happened to be placed late in

the sequence. However, if the prespecified testing sequence accurately reflects the clinical trial's outcome, these procedures can offer power advantages compared to procedures with a data-driven hypothesis ordering.

### 5.4.1 Fixed-sequence procedure

Suppose that there is a natural ordering among the null hypotheses  $H_1, \dots, H_m$ , and the order in which the testing is performed is fixed. (This order normally reflects the clinical importance of the multiple analyses.) The fixed-sequence procedure begins testing the first hypothesis,  $H_1$ , and each test is carried out without a multiplicity adjustment as long as significant results are observed in all preceding tests. The hypothesis  $H_i$  is then rejected at the  $i$ th step if

$$p_j \leq \alpha, j = 1, \dots, i.$$

Otherwise, testing stops, and the remaining hypotheses are accepted without testing. The fixed-sequence procedure controls the FWER because the testing of each hypothesis is conditional upon rejecting all hypotheses earlier in the sequence.

The fixed-sequence testing procedure can be used in a wide variety of multiplicity problems with *a priori* ordered hypotheses. Problems of this kind arise, for example, in clinical trials when longitudinal measurements are analyzed in a sequential manner to identify the onset of therapeutic effect or study its duration.

**EXAMPLE: Case study 2 (Allergen-induced asthma trial)**

Consider a trial designed to assess the efficacy profile of a bronchodilator. Twenty patients with mild asthma were enrolled in this trial and were randomly assigned to receive either an experimental treatment or placebo (10 patients in each treatment group). Patients were given a dose of the treatment and then asked to inhale an allergen to induce bronchoconstriction. Spirometry measurements were taken every 15 minutes for the first hour and every hour up to 3 hours to measure the forced expiratory volume in one second (FEV1). To assess how the treatment attenuated the allergen-induced bronchoconstriction, the FEV1 curves were constructed by averaging the FEV1 values at each time point in the placebo and treated groups. The collected FEV1 data are summarized in Table 5.6.

**TABLE 5.6 Reduction in FEV1 measurements from baseline by time after the allergen challenge (L) in Case study 2**

Time (hours)	Experimental treatment			Placebo		
	n	Mean	SD	n	Mean	SD
0.25	10	0.58	0.29	10	0.71	0.35
0.5	10	0.62	0.31	10	0.88	0.33
0.75	10	0.51	0.33	10	0.73	0.36
1	10	0.34	0.27	10	0.68	0.29
2	10	-0.06	0.22	10	0.37	0.25
3	10	0.05	0.23	10	0.43	0.28

A very important indicator of therapeutic effect is the time to the onset of action, that is, the first time point at which a clinically and statistically significant separation between the FEV1 curves is observed. Since the time points at which spirometry measurements are taken are naturally ordered, the onset of action analyses can be performed using fixed-sequence testing methods. The inferences are performed without an adjustment for multiplicity and without modeling the longitudinal correlation. Thus, the resulting individual tests are more powerful compared to MTPs introduced in Sections 5.2 and 5.3.

Program 5.7 computes and plots the mean treatment differences in FEV1 changes, the associated lower 95% confidence limits, and one-sided *p*-values for the treatment effect from two-sample *t*-tests at each of the 6 time points when the spirometry measurements were taken.

### **PROGRAM 5.7 Treatment comparisons in Case study 2**

```

data fev1;
    input time n1 mean1 sd1 n2 mean2 sd2;
    datalines;
    0.25 10 0.58 0.29 10 0.71 0.35
    0.5 10 0.62 0.31 10 0.88 0.33
    0.75 10 0.51 0.33 10 0.73 0.36
    1 10 0.34 0.27 10 0.68 0.29
    2 10 -0.06 0.22 10 0.37 0.25
    3 10 0.05 0.23 10 0.43 0.28
    ;
run;

data summary;
    set fev1;
    meandif=mean2-mean1;
    se=sqrt((1/n1+1/n2)*(sd1*sd1+sd2*sd2)/2);
    t=meandif/se;
    p=1-probt(t,n1+n2-2);
    lower=meandif-tinv(0.95,n1+n2-2)*se;
run;

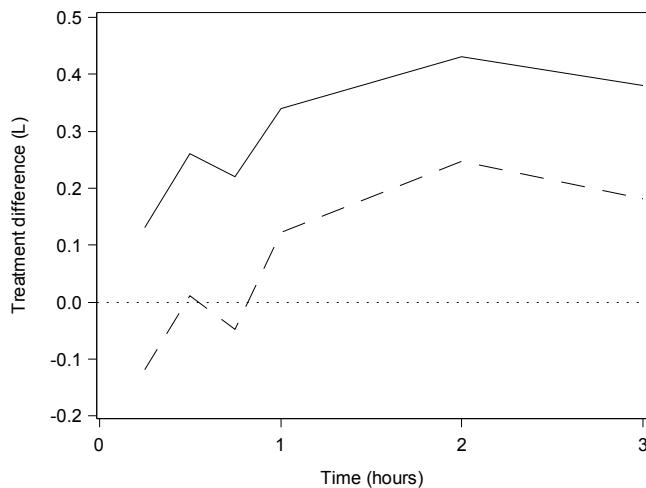
proc sgplot data=summary noautolegend;
    series x=time y=meandif /
        lineattrs=(pattern=solid color=black thickness=1);
    series x=time y=lower /
        lineattrs=(pattern=dash color=black thickness=1);
    refline 0 / lineattrs=(pattern=dot color=black thickness=1);
    yaxis label="Treatment difference (L)" values=(-0.2 to 0.5 by 0.1)
        labelattrs=(family="Arial" size=9pt);
    xaxis label="Time (hours)" values=(0 to 3 by 1)
        labelattrs=(family="Arial" size=9pt);
run;

proc sgplot data=summary noautolegend;
    series x=time y=p /
        lineattrs=(pattern=solid color=black thickness=1);
    refline 0.025 / lineattrs=(pattern=dot color=black thickness=1);
    yaxis label="One-sided p-value" values=(0 to 0.2 by 0.05)
        labelattrs=(family="Arial" size=9pt);
    xaxis label="Time (hours)" values=(0 to 3 by 1)
        labelattrs=(family="Arial" size=9pt);
run;

```

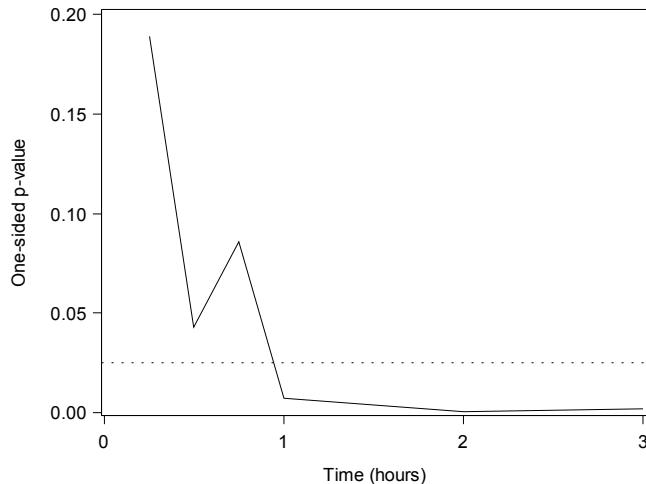
The result of Program 5.7 is shown in Figures 5.7 and 5.8. When reviewing the one-sided confidence limits and raw *p*-values in Figure 5.7, it is tempting to start with the first spirometry and stop testing as soon as a statistically significant mean difference is observed (30 minutes after the allergen challenge). However, this approach does not control the FWER. To protect the FWER, fixed-sequence

**Figure 5.7**  
Case study 2



*Treatment comparisons in Case study 2. Mean treatment difference (---) and lower confidence limit (- - -) by time.*

**Figure 5.8**  
Case study 2



*Treatment comparisons in Case study 2. One-sided raw p-values by time.*

testing should be performed in a sequentially rejective fashion. This means that each subsequent hypothesis is tested only if all previously tested hypotheses were rejected. This can be achieved if we examine the treatment differences beginning with the last spirometry and work backwards. With the correct approach, the lower 95% confidence limit includes zero in Figure 5.7 (or, equivalently, a one-sided *p*-value exceeds the 0.025 threshold in Figure 5.8) for the first time, 45 minutes after the allergen inhalation. As a result, a statistically significant separation between the mean FEV<sub>1</sub> values in the two groups occur 1 hour after the allergen challenge.

This example illustrates the importance of “monotonicity” assumptions in fixed-sequence testing. Fixed-sequence testing methods perform best when the magnitude of the treatment effect can be assumed to change monotonically with respect to time or dose. When the assumption is not met, fixed-sequence procedures are prone to producing spurious results. Coming back to the data summarized in Table 5.6, suppose that the mean difference in FEV<sub>1</sub> changes between the experi-

mental treatment and placebo is very small at the last spirometry. If the associated  $p$ -value is not significant, we cannot determine the onset of therapeutic effect despite the fact that the experimental treatment separated from placebo at several time points.

### 5.4.2 Fallback procedure

The fallback procedure serves as an attractive alternative to the fixed-sequence procedure (Wiens, 2003; Wiens and Dmitrienko, 2005). This procedure implements a more flexible way of handling multiplicity in problems with a prespecified ordering of the hypotheses in a multiplicity problem.

Suppose that the hypotheses  $H_1, \dots, H_m$  are ordered and allocate the overall error rate  $\alpha$  among the hypotheses according to their weights  $w_1, \dots, w_m$  (the weights are non-negative and add up to 1). Specifically, the error rate assigned to  $H_i$  is equal to  $\alpha w_i$ ,  $i = 1, \dots, m$ . The fallback procedure is carried out using the following algorithm that relies on the pre-specified hypothesis ordering:

- Step 1. Test  $H_1$  at  $\alpha_1 = \alpha w_1$ . If  $p_1 \leq \alpha_1$ , reject this hypothesis; otherwise, accept it. Go to the next step.
- Steps  $i = 2, \dots, m - 1$ . Test  $H_i$  at  $\alpha_i = \alpha_{i-1} + \alpha w_i$  if  $H_{i-1}$  is rejected and at  $\alpha_i = \alpha w_i$  if  $H_{i-1}$  is accepted. If  $p_i \leq \alpha_i$ , reject  $H_i$ ; otherwise, accept it. Go to the next step.
- Step  $m$ . Test  $H_m$  at  $\alpha_m = \alpha_{m-1} + \alpha w_m$  if  $H_{m-1}$  is rejected and at  $\alpha_m = \alpha w_m$  if  $H_{m-1}$  is accepted. If  $p_m \leq \alpha_m$ , reject  $H_m$ ; otherwise, accept it.

It is instructive to compare the fallback and fixed-sequence procedures. First, the fallback procedure simplifies to the fixed-sequence procedure when  $w_1 = 1$  and  $w_2 = \dots = w_m = 0$ . If  $w_1 < 1$ , the fallback procedure first tests the primary hypothesis  $H_1$  at a lower significance level compared with the fixed-sequence procedure, i.e.,  $H_1$  is rejected if  $p_1 \leq \alpha w_1$ . This enables a fallback strategy for the subsequent hypotheses because, unlike the fixed-sequence procedure, the fallback procedure can continue to the next hypothesis in the sequence even if the current test is nonsignificant. For example, if  $H_1$  is not rejected,  $H_2$  can still be tested at the significance level  $\alpha w_2$ . On the other hand, if  $H_1$  is rejected, its error rate is carried over to  $H_2$ , which can be tested at a higher significance level, namely,  $\alpha(w_1 + w_2)$ . The same approach is followed to test each hypothesis in the prespecified sequence.

The dose-finding hypertension trial from Case study 1 will be used to illustrate the fallback procedure. As a quick reminder, there are three dose-placebo comparisons performed in this trial with the corresponding hypotheses denoted by  $H_H$  (high dose versus placebo),  $H_M$  (medium dose versus placebo), and  $H_L$  (low dose versus placebo) tested at the two-sided  $\alpha = 0.05$ . If a positive dose-response relationship is anticipated, it is most sensible to order the hypotheses by their expected effect size, i.e., test  $H_H$  first, followed by  $H_M$ , and then  $H_L$ . Further, the sponsor can choose to define unequal hypothesis weights to improve the probability of success at the higher doses, e.g.,

$$w_H = 1/2, \quad w_M = 1/3, \quad w_L = 1/6.$$

Selection of hypothesis weights in this setting is discussed in Dmitrienko and D'Agostino (2013). The initial significance levels assigned to  $H_H$ ,  $H_M$ , and  $H_L$  are given by  $\alpha/2$ ,  $\alpha/3$  and  $\alpha/6$ , respectively.

The testing algorithm used by the fallback procedure can be applied to the two-sided  $p$ -values in Scenario 1 of the hypertension trial example:

$$p_H = 0.015, \quad p_M = 0.0167, \quad p_L = 0.047.$$

The testing algorithm starts with  $H_H$  and rejects it since  $p_H \leq \alpha_1 = \alpha/2 = 0.025$ . After this rejection, the error rate from  $H_H$  can be carried over to  $H_M$ . This hypothesis is tested in Step 2 at the significance level

$$\alpha_2 = \alpha_1 + \alpha/3 = \alpha/2 + \alpha/3 = 5\alpha/6 = 0.0417.$$

Because  $p_M \leq 0.0417$ ,  $H_M$  is rejected and testing continues to Step 3. At this final step,  $H_L$  is tested at

$$\alpha_3 = \alpha_2 + \alpha/6 = 5\alpha/6 + \alpha/6 = \alpha = 0.05$$

and  $H_L$  is rejected because  $p_L \leq 0.05$ . Note that even if  $H_M$  was not rejected in Step 2,  $H_L$  would still be tested but at a lower significance level, namely, at the pre-assigned level of  $\alpha/6 = 0.008$ .

Program 5.8 uses the `%PvalProc` macro to calculate adjusted  $p$ -values using Scenario 1 of the hypertension trial example for the fixed-sequence and fallback procedures. The calculation is performed using the decision matrix algorithm described in Section 5.3.2 since these procedures can also be written as closed testing procedures.

#### PROGRAM 5.8 Fixed-sequence and fallback procedures in Case study 1

```

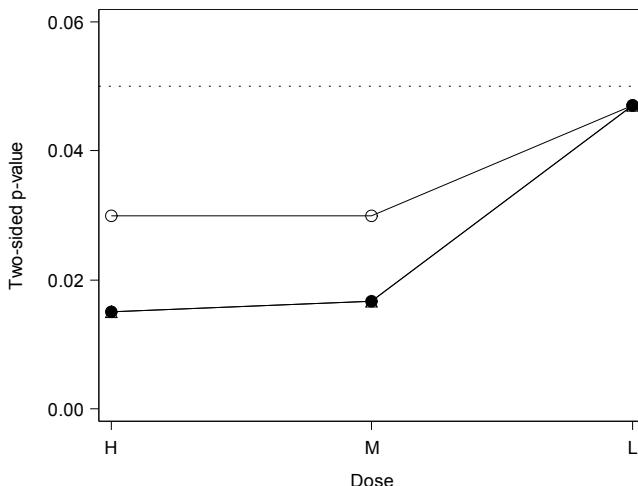
data antihyp1;
    input raw_p weight;
    datalines;
    0.0150 0.5000
    0.0167 0.3333
    0.0470 0.1667
run;

%PvalProc(in=antihyp1,out=adjp);

data adjp;
    set adjp;
    format dose $1.;
    if test=1 then dose="H";
    if test=2 then dose="M";
    if test=3 then dose="L";
run;

proc sgplot data=adjp noautolegend;
    series x=dose y=raw /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=dose y=raw /
        markerattrs=(symbol=circlefilled color=black size=2pct);
    series x=dose y=fallback /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=dose y=fallback /
        markerattrs=(symbol=circle color=black size=2pct);
    series x=dose y=fixedseq /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=dose y=fixedseq /
        markerattrs=(symbol=triangle color=black size=2pct);
    refline 0.05 / lineattrs=(pattern=dot color=black thickness=1);
    yaxis label="Two-sided p-value" values=(0 to 0.06 by 0.02)
        labelattrs=(family="Arial" size=9pt);
    xaxis label="Dose" labelattrs=(family="Arial" size=9pt);
run;
```

The result of Program 5.8 is shown in Figure 5.9. We can see from this figure that both procedures reject all three null hypotheses. The adjusted  $p$ -values calculated based on the fixed-sequence procedure for the High dose and Low dose are smaller compared to those obtained from the fallback procedure (in fact, they are exactly equal to the raw  $p$ -values). This is because the fixed-sequence procedure tests each null hypothesis at the full  $\alpha$  level as long as no hypothesis is accepted in the prespecified sequence. If one hypothesis fails to be rejected, all subsequent null hypotheses are automatically accepted. By contrast, the fallback procedure does not use the full  $\alpha$  level at each step so that all hypotheses are tested even if some hypotheses are accepted earlier in the testing sequence.



**Figure 5.9**  
Case study 1

*Fixed-sequence and fallback procedures. Raw p-value (dot), fallback adjusted p-value (circle), and fixed-sequence adjusted p-value (triangle). Note that the adjusted p-values based on the fixed-sequence procedure are equal to the raw p-values in this particular example.*

### 5.4.3 Chain procedure

The class of nonparametric procedures known as *chain procedures* provides an extension of the fixed-sequence and fallback procedures (Bretz et al., 2009, 2011a; Millen and Dmitrienko, 2011). Chain procedures use flexible decision rules that can be visualized using diagrams that are easy to communicate to clinical development teams. Although this broad class also includes procedures that rely on a data-driven hypothesis ordering (e.g., Holm procedure) this section discusses chain procedures based on *a priori* ordered hypotheses. Note that chain procedures can also be used to implement Bonferroni-based gatekeeping procedures introduced in Section 5.6.

Chain procedures are similar to the fallback procedure in the sense that weights are pre-assigned to each hypothesis in a multiplicity problem. These weights determine the initial significance levels for these hypotheses. The main difference with the fallback procedure lies in the fact that chain procedures support more flexible  $\alpha$  propagation rules, e.g., after each rejection, the error rate can be transferred simultaneously to several hypotheses. By using flexible  $\alpha$  propagation rules, the trial's sponsor can set up more complex decision trees that go beyond a single sequence of hypotheses.

For example, suppose that in the dose-finding hypertension trial from Case study 1, data available from the Phase II trial suggest that the high dose is more

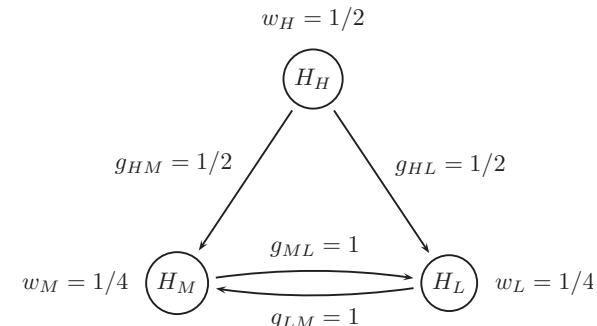
efficacious than the medium and low doses, but that the efficacy of the medium dose is likely to be very similar to the low dose. In this scenario, the fixed-sequence and fallback procedures might be suboptimal because there is no clear hierarchical testing order between the low and medium doses. Instead, the sponsor can construct a chain procedure with a greater weight assigned to  $H_H$  (high dose versus placebo) and split the remaining weight equally between  $H_M$  (medium dose versus placebo) and  $H_L$  (low dose versus placebo).

The resulting testing scheme is illustrated in Figure 5.10. It is based on the initial hypothesis weights  $w_H = 1/2$ ,  $w_M = 1/4$ , and  $w_L = 1/4$ , which reflect the relative importance of the three hypotheses. The resulting initial significance levels for the three hypotheses are then given by  $\alpha/2$ ,  $\alpha/4$ , and  $\alpha/4$ . The transition parameters  $g_{HM}$ ,  $g_{HL}$ ,  $g_{ML}$ , and  $g_{LM}$  determine how  $\alpha$  is propagated in case a hypothesis is rejected. More specifically:

- The error rate released after the rejection of  $H_H$  is split between  $H_M$  and  $H_L$ .
- If  $H_M$  is rejected, the error rate is transferred to  $H_L$ . Similarly, if  $H_L$  is rejected, the error rate is transferred to  $H_M$ .

Every time a hypothesis is rejected, the hypothesis weights and transition parameters are updated following the algorithm given in Bretz et al. (2009). See the chapter Appendix for more information.

**Figure 5.10**  
Case study 1



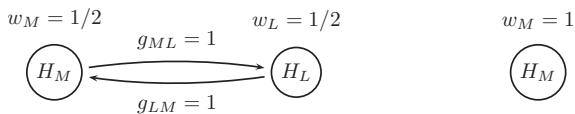
*Visual summary of the chain procedure in Case study 1.*

In Scenario 4 of the hypertension trial example, the following raw  $p$ -values are observed:

$$p_H = 0.017, p_M = 0.026, p_L = 0.022.$$

Beginning with the high dose,  $H_H$  can be rejected at the two-sided  $\alpha = 0.05$  since  $p_H \leq \alpha/2 = 0.025$ . It follows that its error rate can be split between  $H_M$  and  $H_L$  based on the prespecified transition parameters  $g_{HM} = g_{HL} = 1/2$ . The weights and transition parameters can then be updated as shown in the left panel of Figure 5.11. The hypothesis  $H_M$  can now be tested at the significance level  $\alpha/2 = 0.025$  but fails to be rejected at this step since  $p_M = 0.026$ . Therefore, no further  $\alpha$  can be transferred to  $H_L$ . The hypothesis  $H_L$  is tested at the same significance level  $\alpha/2$  and is successfully rejected because  $p_L = 0.022 \leq 0.025$ . An important feature of the decision rules used in this chain procedure is that, unlike the simple fallback procedure, the chain procedure enables a symmetric error rate exchange between the hypotheses  $H_M$  and  $H_L$ . Since the chain procedure rejects  $H_L$ , the error rate released after this rejection is carried forward to  $H_M$ , and its weight  $w_M$  is updated, as shown on the right panel of Figure 5.11. As a result,  $H_M$  can be tested at the full  $\alpha$  level and is finally rejected since  $p_M = 0.026 \leq 0.05$ .

**Figure 5.11**  
Case study 1



Visual summary of the updated chain procedure in Case study 1. Left panel: Updated chain procedure after rejecting  $H_H$ . Right panel: Updated chain procedure after rejecting both  $H_H$  and  $H_L$ .

Program 5.9 computes adjusted  $p$ -values based on the chain procedure using the %Chain macro. This macro uses the decision matrix algorithm described in Section 5.3.2 and computes adjusted  $p$ -values. (Note that chain procedures can also be defined using the closure principle.) The macro requires the following two parameters:

- **in** is the name of the data set with raw  $p$ -values and procedure parameters. This data set must contain one row per null hypothesis and the following variables:
  - **raw\_p**: raw  $p$ -value.
  - **weight**: hypothesis weight.
  - **g1** through **gn**: transition parameters. (Here, **n** is the number of null hypotheses.)
 The null hypotheses in this data set are assumed to be ordered from first to last.
- **out** is the name of the data set with the adjusted  $p$ -values.

### PROGRAM 5.9 Chain procedure in Case study 1

```

data antihyp4;
  input raw_p weight g1-g3;
  datalines;
  0.017 0.50 0.0 0.5 0.5
  0.026 0.25 0.0 0.0 1.0
  0.022 0.25 0.0 1.0 0.0
run;

%Chain(in=antihyp4,out=adjp);

data adjp;
  set adjp;
  format dose $1.;
  if test=1 then dose="H";
  if test=2 then dose="M";
  if test=3 then dose="L";
run;

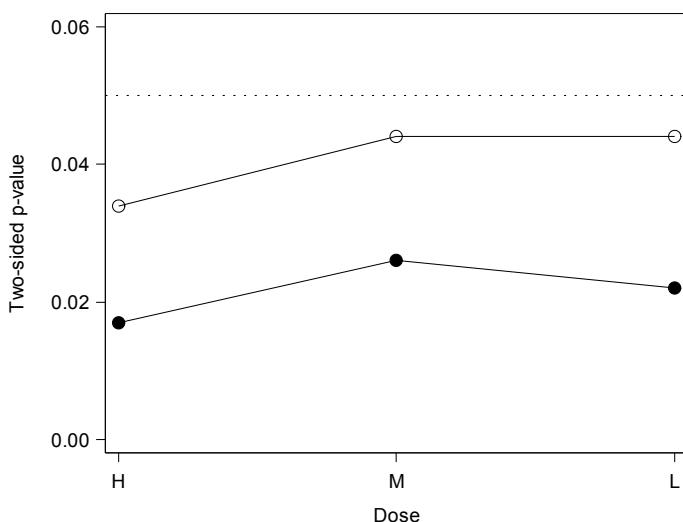
proc sgplot data=adjp noautolegend;
  series x=dose y=raw /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=dose y=raw /
    markerattrs=(symbol=circlefilled color=black size=2pct);
  series x=dose y=chain /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=dose y=chain /
    markerattrs=(symbol=circle color=black size=2pct);
  refline 0.05 / lineattrs=(pattern=dot color=black thickness=1);
  yaxis label="Two-sided p-value" values=(0 to 0.06 by 0.02)
    labelattrs=(family="Arial" size=9pt);
  xaxis label="Dose" labelattrs=(family="Arial" size=9pt);
run;

```

As shown in Program 5.9, the input data set (`antihyp4` data set) contains the raw  $p$ -values for the three null hypotheses of no effect as well as the initial hypothesis weights and transition parameters of the chain procedure define above.

The result of Program 5.9 is shown in Figure 5.12. Figure 5.12 shows that all three adjusted  $p$ -values calculated based on the chain procedure are less than  $\alpha = 0.05$ . This implies that all three null hypotheses are rejected. It is interesting to notice that the adjusted  $p$ -values corresponding to the medium and low doses are equal. As explained above, this reflects the fact that  $H_M$  can only be rejected after  $H_L$  is first rejected and its error rate is transferred to  $H_M$ .

**Figure 5.12**  
Case study 1



Chain procedure. Raw  $p$ -value (dot) and chain adjusted  $p$ -value (circle).

#### 5.4.4 Summary

Multiple testing procedures introduced in this section rely on a prespecified hypothesis ordering and provide an alternative to data-driven hypothesis testing when the null hypotheses of interest are naturally ordered. In the presence of an *a priori* ordering, we can, for example, test the individual hypotheses using the fixed-sequence procedure in a sequentially rejective fashion without any adjustment for multiplicity. Each subsequent hypothesis is tested only if all previously tested hypotheses were rejected.

The fixed-sequence procedure is known to perform best when the magnitude of treatment effect can be assumed to change monotonically with respect to time or dose. As an example, it was demonstrated in this section how to perform fixed-sequence inferences to identify the onset of therapeutic effect in an asthma study. When this monotonicity assumption is not met, fixed-sequence procedures are likely to produce spurious results.

The fallback procedure was proposed as a more flexible alternative to the fixed-sequence procedure. This procedure also tests hypotheses following a prespecified ordering, but it does not use the full available  $\alpha$ -level at each step, which enables a fallback strategy in case a null hypothesis is failed to be rejected.

It was demonstrated that the fixed-sequence and fallback procedures are, in fact, examples of a broad class of MTPs, namely, chain procedures. These procedures allow trial sponsors to set up custom decision rules with flexible hypothesis weight allocation and  $\alpha$  propagation rules. It should be noted that the chain procedures described in this section are based on a nonparametric approach where no assumptions

are made regarding the joint distribution of the hypothesis test statistics. Refer to Bretz et al. (2011a) for a description of chain procedures based on semiparametric and parametric approaches (see Section 5.5) and their implementation in SAS.

The fixed-sequence and fallback procedures can be implemented using either the %PvalProc or %Chain macros. General chain procedures are supported by the %Chain macro.

## 5.5 Parametric procedures

---

In the previous sections, two types of MTPs were introduced: nonparametric procedures (e.g., Bonferroni and fixed-sequence), and semiparametric procedures (e.g., Hochberg or Hommel). The key feature of nonparametric procedures is that they impose no distributional assumptions. But semiparametric procedures rely on fairly flexible distributional assumptions. When the joint distribution of the hypothesis test statistics is fully specified (e.g., in dose-finding clinical trials with normally distributed outcomes), the trial's sponsor can consider parametric procedures that can offer a power advantage over nonparametric and semiparametric procedures.

As opposed to  $p$ -value based procedures, parametric procedures are formulated on the basis of specific distributional models, and their critical values are computed directly from the joint distribution of the test statistics. The most well known parametric procedure is the Dunnett procedure (Dunnett, 1955), developed for problems with multiple dose-control comparisons. The Dunnett single-step procedure and its extensions to stepwise procedures are described in this section.

The following setting will be used throughout this section. Consider a one-way ANOVA model frequently used in clinical trials with several dose-control comparisons. Assume that  $m$  doses are tested versus a control (e.g., a placebo) and consider a balanced design with  $n$  patients per trial arm. This corresponds to the following ANOVA model:

$$y_{ij} = \mu_i + \varepsilon_{ij}.$$

Here  $i$  is the trial arm index,  $i = 0, \dots, m$ , and  $i = 0$  denotes the control arm. Further,  $j$  is the patient index;  $j = 1, \dots, n$ ,  $y_{ij}$  denotes the response;  $\mu_i$  is the mean effect; and  $\varepsilon_{ij}$  is the error term. The error term is assumed to be normally distributed with mean 0 and standard deviation  $\sigma$ .

Assuming that a larger value of the response variable indicates a beneficial effect, the null hypotheses in this multiplicity problem are defined as follows:

$$H_i : \theta_i \leq 0,$$

where  $\theta_i = \mu_i - \mu_0$  is the mean treatment difference,  $i = 1, \dots, m$ . The hypotheses will be tested using the following test statistics:

$$t_i = \frac{\bar{y}_i - \bar{y}_0}{s\sqrt{2/n}}, \quad i = 1, \dots, m,$$

where  $s$  is the pooled sample standard deviation.

The following example will be used to illustrate the Dunnett procedures introduced below.

### **EXAMPLE: Case study 3 (Dose-finding dyslipidemia trial)**

Consider a dose-finding trial in patients with dyslipidemia. The trial is conducted to compare the effect of four doses of an experimental treatment, labeled D1 (lowest dose) through D4 (highest dose), to that of a placebo. The primary efficacy endpoint is based on the mean increase in HDL cholesterol at 12 weeks. The sample size in each treatment group is 26 patients.

Program 5.10 uses PROC MULTTEST and PROC MIXED to summarize the mean HDL cholesterol at 12 weeks by treatment group (Figure 5.13) and the related mean increase compared to placebo for each dose (Output 5.10).

### PROGRAM 5.10 Data set in Case study 3

```

data dftrial(drop=dose0-dose4);
    array dose[*] dose0-dose4;
    input dose0-dose4 @@;
    do group=0 to 4;
        hdl=dose[group+1];
        output;
    end;
    datalines;
41.0 44.6 38.8 48.7 39.4 43.2 42.8 49.8 44.5 41.4 44.0 43.5 39.3
44.9 38.3 38.9 43.7 39.3 45.0 42.0 38.8 36.5 42.2 44.5 40.0 34.4
42.7 35.4 40.4 46.7 39.4 42.8 39.9 44.6 46.6 44.5 41.4 39.3 44.9
47.0 36.8 36.0 45.7 39.3 47.4 43.2 42.0 38.1 42.2 37.5 38.8 41.7
43.2 39.3 45.0 41.0 44.0 48.1 43.4 42.2 38.9 40.6 41.6 44.6 46.2
34.8 45.6 51.4 35.5 45.8 36.8 46.0 39.0 49.9 37.6 39.4 42.3 43.5
41.3 48.6 45.2 44.7 40.7 49.2 43.2 37.7 39.1 42.2 41.3 42.6 38.4
38.5 38.9 45.3 38.8 35.2 44.7 39.8 34.0 42.7 43.1 41.7 42.8 44.5
43.0 38.5 43.6 41.2 39.0 44.2 32.0 33.8 39.2 48.1 49.9 39.0 41.9
47.7 40.7 45.4 37.4 38.1 43.7 50.2 43.7 45.7 37.9 41.0 36.9 45.0
run;

proc multtest data=dftrial;
    class group;
    test mean(hdl);
    ods output continuous=doseresp;
run;

proc sgplot data=doseresp noautolegend;
    series x=group y=mean /
        lineattrs=(pattern=solid color=black thickness=1);
    scatter x=group y=mean /
        markerattrs=(symbol=circlefilled color=black size=2pct);
    yaxis label="Mean HDL cholesterol (mg/dl)" values=(35 to 45 by 5)
        labelattrs=(family="Arial" size=9pt);
    xaxis label="Treatment group" labelattrs=(family="Arial" size=9pt);
run;

ods output Diffs=LSdiffs ;
proc mixed data = dftrial;
    class group;
    model hdl = group;
    lsmeans group / pdiff=controlu;
run;

proc print data=LSdiffs noobs label;
    var group estimate StdErr DF tValue Probt;
run;

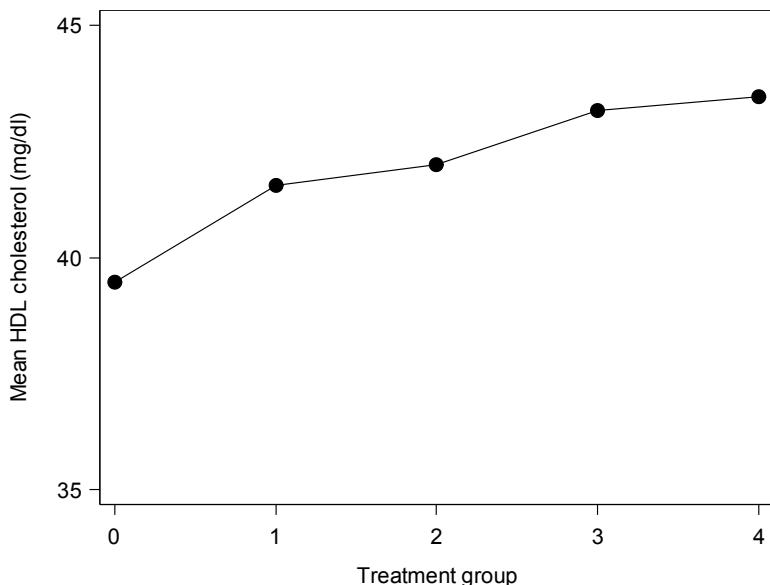
```

**Output from Program 5.10** Differences of Least Squares Means

group	Estimate	Standard		t Value	Pr > t
		Error	DF		
1	2.0808	1.0176	125	2.04	0.0215
2	2.5269	1.0176	125	2.48	0.0072
3	3.6962	1.0176	125	3.63	0.0002
4	4.0038	1.0176	125	3.93	<.0001

Output 5.10 summarizes the mean increase in HDL cholesterol at 12 weeks for each dose compared to placebo; associated standard error; degrees of freedom (df); *t*-statistics; and one-sided raw *p*-values. Looking at Figure 5.13 and Output 5.10, each dose of the experimental treatment seems to provide a beneficial increase compared to placebo in HDL cholesterol. It is then desirable to formally test the null hypotheses  $H_1$ ,  $H_2$ ,  $H_3$ , and  $H_4$  defined above while controlling the FWER in the strong sense.

**Figure 5.13**  
Case study 3



Mean HDL cholesterol (mg/dl) at 12 weeks by trial arm.

### 5.5.1 Single-step Dunnett procedure

The single-step Dunnett procedure can be thought of as the parametric counterpart of the Bonferroni procedure. Unlike the Bonferroni procedure, which does not depend on any distributional assumptions, the Dunnett procedure is based on the joint distribution of the test statistics associated with the hypotheses of interest. Thus, it accounts for the correlations among the test statistics. The use of the Dunnett procedure leads to more powerful inferences compared to the Bonferroni procedure.

Under the setting introduced at the beginning of this section, the  $m$  test statistics  $t_i$  follow a fully specified multivariate  $t$  distribution, and, given a balanced design, it is easy to show that the test statistics are equally correlated with the common correlation coefficient of 0.5. Using this fact, Dunnett (1955) proposed to define the common one-sided critical value for the hypothesis test statistics  $t_1, \dots, t_m$ . This critical value is denoted by  $d_\alpha(m, \nu)$  and is found as the  $(1 - \alpha)$ -quantile of the distribution of the maximum of  $t$ -distributed random variables with  $\nu = (m+1)(n-1)$  degrees of freedom, i.e.,

$$d_\alpha(m, \nu) = F^{-1}(1 - \alpha | m, \nu).$$

Here  $F(x|m, \nu)$  is the cumulative distribution function of the one-sided Dunnett distribution, i.e.,

$$F(x|m, \nu) = P\{\max(T_1, \dots, T_m) \leq x\},$$

and  $T_1, \dots, T_m$  are the random variables that follow the same distribution as the hypothesis test statistics  $t_1, \dots, t_m$  under the global null hypothesis of no effect across the  $m$  doses. The Dunnett procedure rejects the null hypothesis  $H_i$  if its test statistics is greater or equal to the common critical value, i.e., if

$$t_i \geq d_\alpha(m, \nu), \quad i = 1, \dots, m.$$

For example, the hypothesis test statistics in Case study 3, calculated using Program 5.10, are equal to

$$t_1 = 2.08, \quad t_2 = 2.53, \quad t_3 = 3.70, \quad t_4 = 4.00.$$

The one-sided, Dunnett-adjusted critical value is given by  $d_\alpha(m, \nu)$  with  $\alpha = 0.025$ ,  $m = 4$ , and  $\nu = (m + 1)(n - 1) = 125$ . It can be calculated using the PROBMC function as illustrated in Program 5.11.

### PROGRAM 5.11 Dunnett-adjusted critical value in Case study 3

```
data crit;
  d=probmc("DUNNETT1", ., 0.975, 125, 4);
run;

proc print data=crit noobs;
  title "Dunnett-adjusted critical value";
run;
```

---

#### Output from Program 5.11

```
Dunnett-adjusted critical value
d
2.47337
```

---

As shown in Output 5.11, the critical value is equal to 2.47. Therefore, the Dunnett procedure concludes that the treatment effect at Doses D2, D3, and D4 are significant.

Adjusted  $p$ -values are commonly used to present the results of Dunnett-based multiplicity adjustments. Program 5.12 uses PROC MIXED to calculate the adjusted  $p$ -values based on the single-step Dunnett and Bonferroni procedures. (Note that this can also be achieved using PROC GLM and PROC GLIMMIX.) The individual patient data are contained in the dftrial data set that was created in Program 5.10.

**PROGRAM 5.12 Single-step Dunnett and Bonferroni procedures in Case study 3**

```

ods output Diffs=LSdiffs ;
proc mixed data = dftrial;
  class group;
  model hdl = group;
  lsmeans group / adjust=dunnett pdiff=controlu;
  lsmeans group / adjust=bon pdiff=controlu;
run;

data dunnett (keep=group raw dunnett);
  set LSdiffs;
  where adjustment="Dunnett";
  rename Probt=raw adjp=dunnett;
run;

data bonf (keep=group bonf);
  set LSdiffs;
  where adjustment="Bonferroni";
  rename adjp=bonf;
run;

data adjp;
  merge dunnett bonf;
  by group;
run;

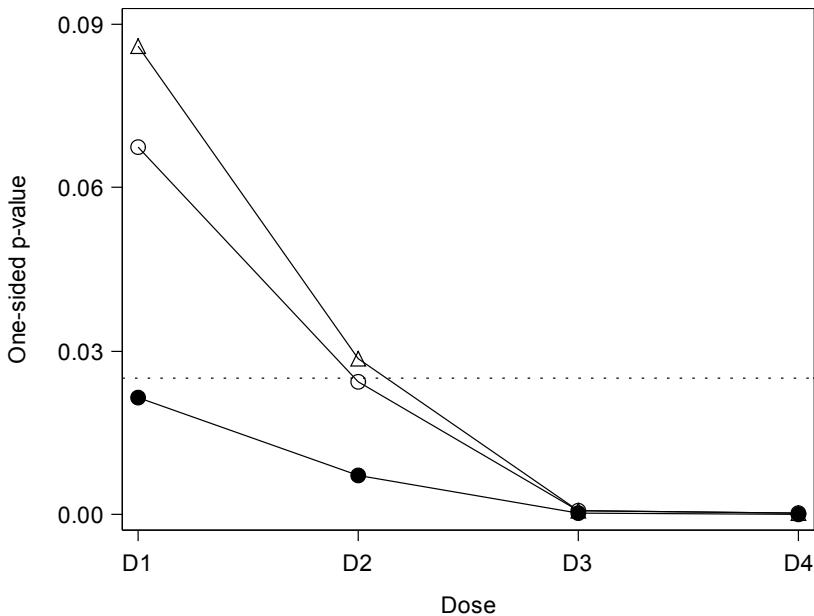
data adjp;
  set adjp;
  format dose $2. ;
  if group=1 then dose="D1";
  if group=2 then dose="D2";
  if group=3 then dose="D3";
  if group=4 then dose="D4";
run;

proc sgplot data=adjp noautolegend;
  series x=dose y=raw /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=dose y=raw /
    markerattrs=(symbol=circlefilled color=black size=2pct);
  series x=dose y=dunnett /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=dose y=dunnett /
    markerattrs=(symbol=circle color=black size=2pct);
  series x=dose y=bonf /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=dose y=bonf /
    markerattrs=(symbol=triangle color=black size=2pct);
  refline 0.025 / lineattrs=(pattern=dot color=black thickness=1);
  yaxis label="One-sided p-value" values=(0 to 0.09 by 0.03)
    labelattrs=(family="Arial" size=9pt);
  xaxis label="Dose" labelattrs=(family="Arial" size=9pt);
run;

```

The result of Program 5.12 is displayed in Figure 5.14. This figure shows that the adjusted  $p$ -values calculated using the single-step Dunnett procedure are smaller than those obtained using the Bonferroni procedure. In addition, the Bonferroni procedure fails to reject the null hypothesis associated with Dose D2. But it can be rejected using the single-step Dunnett procedure. In fact, it can be shown that the single-step Dunnett procedure is uniformly more powerful than the Bonferroni procedure.

**Figure 5.14**  
Case study 3



Single-step Dunnett and Bonferroni procedures. Raw  $p$ -value (dot), single-step Dunnett adjusted  $p$ -value (circle), and Bonferroni adjusted  $p$ -value (triangle).

### 5.5.2 Stepwise Dunnett procedures

The single-step Dunnett procedure defined in Section 5.5.1 is straightforward to apply as it uses a common critical value to test all null hypotheses. As discussed in Section 5.3, single-step procedures can be improved by applying the closure principle which results in more powerful stepwise procedures based on a data-driven testing sequence. For example, the Holm procedure uses a step-down testing algorithm and is superior to the single-step Bonferroni procedure. Similarly, stepwise extensions of the single-step Dunnett procedure can be defined as illustrated below.

Recall from Section 5.3 that stepwise procedures such as the Holm were defined using ordered  $p$ -values. In order to define the parametric stepwise procedures, it is more convenient to work with ordered test statistics  $t_{(1)} > \dots > t_{(m)}$  and associated ordered hypotheses  $H_{(1)}, \dots, H_{(m)}$ .

#### Step-down Dunnett procedure

The step-down Dunnett procedure (Naik, 1975; Marcus et al., 1976; Dunnett and Tamhane, 1991) is a sequentially rejective procedure that begins with  $H_{(1)}$ , i.e., the

null hypothesis associated with the largest test statistic or, equivalently, the smallest  $p$ -value. If this hypothesis cannot be rejected, testing stops and the remaining hypotheses are automatically accepted. Otherwise, the next hypothesis in the sequence,  $H_{(2)}$ , is tested and the testing algorithm continues in the same sequentially rejective manner. This algorithm is similar to the one used in the Holm procedure, and, in fact, the step-down Dunnett procedure can be seen as a parametric extension of the Holm procedure.

To define a step-down Dunnett procedure, let  $d_\alpha(i, \nu)$ ,  $i = 1, \dots, m$ , denote the  $(1 - \alpha)$ -quantile of the  $i$ -variate  $t$  distribution with  $\nu = (m + 1)(n - 1)$  degrees of freedom. Using this notation, the following testing algorithm is used in the step-down Dunnett procedure:

- Step 1. Reject  $H_{(1)}$  if  $t_{(1)} \geq c_1$ , where  $c_1 = d_\alpha(m, \nu)$ . If this hypothesis is rejected, proceed to Step 2. Otherwise, accept all hypotheses and stop.
- Steps  $i = 2, \dots, m - 1$ . Reject  $H_{(i)}$  if  $t_{(i)} \geq c_i$ , where  $c_i = d_\alpha(m - i + 1, \nu)$ . If this hypothesis is rejected, proceed to Step  $i + 1$ . Otherwise, accept all remaining hypotheses and stop.
- Step  $m$ . Reject  $H_{(m)}$  if  $t_{(m)} \geq c_m$ , where  $c_m = d_\alpha(1, \nu)$ . Otherwise, accept  $H_{(m)}$ .

It is helpful to note that the step-down Dunnett procedure uses the critical value used in the single-step Dunnett procedure at the first step ( $c_1$ ). However, the other null hypotheses are tested using successively lower critical values, namely,

$$c_1 > c_2 > \dots > c_m.$$

This shows that the step-down procedure rejects as many, and possibly more, null hypotheses compared to the single-step Dunnett procedure. Further, it is easy to prove that the step-down Dunnett procedure is also uniformly more powerful than the Holm procedure.

Considering the dyslipidemia trial from Case study 3, the test statistics were calculated using Program 5.10 and are ordered as follows:

$$t_{(1)} = t_4 = 4.00, t_{(2)} = t_3 = 3.70, t_{(3)} = t_2 = 2.53, t_{(4)} = t_1 = 2.08.$$

The critical values used in Steps 1 to 4 of the step-down Dunnett algorithm are equal to  $c_1 = 2.47$ ,  $c_2 = 2.38$ ,  $c_3 = 2.24$ , and  $c_4 = 1.98$ . (These values can be calculated using the PROBMC function.) Comparing each of the ordered statistics to the corresponding critical value using the step-down algorithm, it can be seen that all null hypotheses can be rejected. Recall that, with the single-step Dunnett procedure,  $H_1$  could not be rejected, which illustrates the advantage of the stepwise extension over its single-step counterpart.

Program 5.13 proposes an implementation of the step-down Dunnett and Holm procedures in Case study 3 using PROC GLIMMIX. This program computes the adjusted  $p$ -values produced by the two procedures. As a reminder, the individual patient data are contained in the `dftrial` data set.

**PROGRAM 5.13 Step-down Dunnett and Holm procedures in Case study 3**

```

ods output Diffss=LSdiffs;
proc glimmix data = dftrial;
  class group;
  model hdl = group;
  lsmeans group / adjust = dunnett stepdown diff = controlu;
  lsmeans group / adjust = bon stepdown diff = controlu;
run;

data stepdun (keep=group raw stepdun);
  set LSdiffs;
  where adjustment="Stepdown Dunnett";
  rename Probt=raw adjp=stepdun;
run;

data holm (keep=group holm);
  set LSdiffs;
  where adjustment="Stepdown Bonferroni";
  rename adjp=holm;
run;

data adjp;
  merge stepdun holm;
  by group;
run;

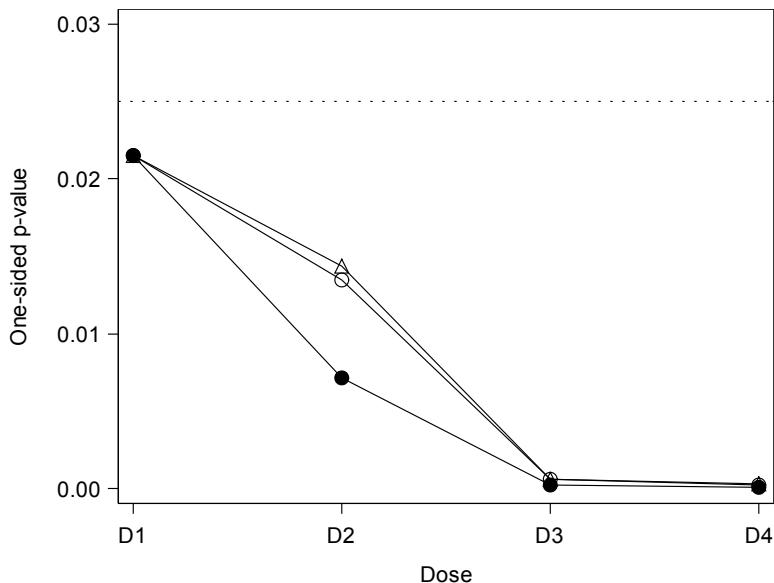
data adjp;
  set adjp;
  format dose $2.;
  if group=1 then dose="D1";
  if group=2 then dose="D2";
  if group=3 then dose="D3";
  if group=4 then dose="D4";
run;

proc sgplot data=adjp noautolegend;
  series x=dose y=raw /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=dose y=raw /
    markerattrs=(symbol=circlefilled color=black size=2pct);
  series x=dose y=stepdun /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=dose y=stepdun /
    markerattrs=(symbol=circle color=black size=2pct);
  series x=dose y=holm /
    lineattrs=(pattern=solid color=black thickness=1);
  scatter x=dose y=holm /
    markerattrs=(symbol=triangle color=black size=2pct);
  refline 0.025 / lineattrs=(pattern=dot color=black thickness=1);
  yaxis label="One-sided p-value" values=(0 to 0.03 by 0.01)
    labelattrs=(family="Arial" size=9pt);
  xaxis label="Dose" labelattrs=(family="Arial" size=9pt);
run;

```

The result of Program 5.13 is shown in Figure 5.15. It is demonstrated in this figure that the step-down Dunnett and Holm procedures both reject all four null hypotheses at a one-sided  $\alpha = 0.025$ . Even though the two procedures lead to the same number of rejections, it can be seen that the adjusted  $p$ -values obtained using the step-down Dunnett procedure are either equal to or more significant than those produced by the Holm procedure.

**Figure 5.15**  
Case study 3



*Step-down Dunnett and Holm procedures. Raw p-value (dot), step-down Dunnett adjusted p-value (circle), and Holm adjusted p-value (triangle).*

### Step-up Dunnett procedure

The step-up Dunnett procedure (Dunnett and Tamhane, 1992) serves as the parametric extension of the Hochberg procedure. It uses a similar step-up algorithm and begins with the null hypothesis corresponding to the least significant test statistic. If this null hypothesis is rejected, testing stops and all remaining null hypotheses are rejected. The step-up Dunnett procedure is uniformly more powerful than the single-step Dunnett procedure as well as the Hochberg procedure. However, the step-up Dunnett procedure does not uniformly dominate the step-down Dunnett procedure in terms of power. In addition, the derivation of the critical values of the step-up Dunnett procedure is a computationally intensive problem, especially in the unbalanced case, and this procedure is rarely used in clinical trials.

### 5.5.3 Summary

The parametric procedures defined in this section rely on the assumption that the joint distribution of the hypothesis test statistics is fully specified. In this setting, parametric procedures provide a power advantage over  $p$ -value based procedures defined in Sections 5.2 and 5.3.

In particular, the single-step Dunnett procedure corresponds to a parametric extension of the Bonferroni procedure. This single-step procedure dominates the Bonferroni procedure in terms of power, and it uses a common critical value to test all null hypotheses. The single-step Dunnett procedure can be implemented using PROC GLM, PROC MIXED, and PROC GLIMMIX.

As shown in Section 5.3, more powerful alternatives to single-step procedures can be built using the closure principle, and the resulting procedures can be used to test hypotheses in a stepwise manner. Similarly, the single-step Dunnett procedure can be extended to set up stepwise parametric procedures. The step-down Dunnett procedure is similar to the Holm procedure in the sense that Dunnett first tests the hypothesis corresponding to the most significant test statistics. The step-down Dunnett procedure is uniformly more powerful than the single-step Dunnett procedure as well as the Holm procedure. The step-down Dunnett procedure can be implemented using PROC GLIMMIX.

A step-up version of the Dunnett procedure can also be defined and is conceptually similar to the Hochberg procedure. The step-up Dunnett procedure is uniformly more powerful than the single-step Dunnett procedure and Hochberg procedures but is not always more powerful than its step-down alternative.

It is important to note that, although a one-way ANOVA model setting was used in this section to illustrate parametric procedures, a parametric approach can be applied broadly to all settings with a known joint distribution of the test statistics. Further, multivariate  $t$  distributions used in the computation of the critical values of the parametric procedures in the Dunnett family can be replaced with multivariate normal distributions. This enables, for example, the use of parametric procedures in clinical trials with binary endpoints.

## 5.6 Gatekeeping procedures

---

This section describes multiple testing procedures used in clinical trials with a complex set of objectives that represent several sources of multiplicity, including multiplicity problems with multiple endpoints, multiple doses, noninferiority/superiority tests, and subgroup analyses. In this setting, null hypotheses are grouped into families that are tested in a sequential manner in the sense that the acceptance or rejection of null hypotheses in a particular family depends on the outcome of the significance tests carried out in the preceding families. In other words, the families of hypotheses examined earlier serve as *gatekeepers*, and we can test hypotheses in the current family only if the preceding gatekeepers were successfully passed.

Gatekeeping procedures serve as the main tool for addressing complex multiplicity issues of this kind. Several types of gatekeeping procedures for hierarchically ordered families of hypotheses have been proposed in the literature, including procedures for problems with serial gatekeepers (Maurer et al., 1995; Bauer et al., 1998; Westfall and Krishen, 2001), parallel gatekeepers (Dmitrienko et al., 2003, 2008), and general gatekeepers (Dmitrienko and Tamhane, 2011, 2013). To efficiently address multiplicity issues in settings with general gatekeepers, Dmitrienko and Tamhane (2011, 2013) introduced the mixture method for building gatekeeping procedures. This method can be used for defining gatekeeping procedures based on the Bonferroni as well as more powerful procedures, e.g., Hommel-based gatekeeping (Brechenmacher et al., 2011). They can be tailored to various clinical trial designs. Other MTPs that can be applied in clinical trials with ordered families include nonparametric (Bonferroni-based) and parametric chain procedures (see Section 5.4.3). A detailed description of these procedures and their key properties can be found in recently published review papers and tutorials, for example, Bretz et al.

(2009, 2011a), Dmitrienko et al. (2013), and Alesh et al. (2014). Commonly used gatekeeping procedures are also described in the FDA's draft guidance document on multiple endpoints in clinical trials (FDA, 2017).

In this section, we will focus on gatekeeping methods based on nonparametric and semiparametric procedures, e.g., Bonferroni, Holm, Hochberg, and Hommel procedures introduced in Sections 5.2 and 5.3. The cases studies provided below will be used to illustrate the standard mixture method originally developed in Dmitrienko and Tamhane (2011, 2013) as well as a modified mixture method that is tailored for problems with multiple sequences of hypotheses (Dmitrienko et al., 2016). Gatekeeping methods based on parametric procedures are discussed in Dmitrienko and Tamhane (2011, 2013).

### 5.6.1 Examples of ordered families of hypotheses in clinical trials

Focusing on multiplicity problems arising in confirmatory clinical trials, null hypotheses of no treatment effect are commonly divided into primary and secondary (FDA, 2017). The primary hypotheses are formulated in terms of the primary trial objectives that are related to the most important features of an experimental treatment. In most confirmatory trials, the primary analyses determine the overall outcome of the trial and provide the basis for the key regulatory claims. Secondary hypotheses might also play an important role in determining the overall outcome, and it is often desirable to present secondary findings in the product label.

The following clinical trial examples present different scenarios where the secondary analyses can potentially lead to additional regulatory claims. In this case, the FWER must be controlled strongly with respect to the primary and secondary families of hypotheses. The first two case studies will be used to construct more basic Bonferroni-based gatekeeping procedures. The third case study will be used to illustrate a more powerful gatekeeping procedure derived from a semiparametric MTP. The last case study will also be used to introduce the modified mixture method and explore the magnitude of power gain compared to the standard mixture method.

**EXAMPLE: Case study 4 (Depression trial)**

Consider a Phase III clinical trial in patients with clinical depression. The efficacy profile of a novel treatment is characterized using several clinical endpoints (primary and key secondary endpoints). These endpoints are defined as follows:

- Primary endpoint is based on the mean improvement from baseline in the 17-item Hamilton Depression Scale (HAMD17 score).
- Key secondary endpoints include the response and remission rates based on the HAMD17 score.

The primary regulatory claim will be based on the primary outcome variable, and the secondary variables will be analyzed to evaluate the evidence for additional regulatory claims. The primary hypothesis serves as a serial gatekeeper in the sense that the secondary hypotheses will be tested only after the primary analysis has yielded a statistically significant result. We wish to develop a gatekeeping strategy for testing the primary and two secondary hypotheses in a manner that preserves the FWER in the strong sense.

**EXAMPLE: Case study 5 (Acute respiratory distress syndrome trial)**

This case study deals with a Phase III clinical trial in patients with acute respiratory distress syndrome (ARDS). The trial is conducted to compare one dose of an

investigational treatment to placebo. The therapeutic benefits of experimental treatments in ARDS trials are commonly measured using the following primary endpoints:

- Number of days a patient is alive and is on mechanical ventilation during the 28-day study period.
- 28-day all-cause mortality.

Either of these two endpoints can be used to make regulatory claims.

Additional regulatory claims can be made with respect to secondary endpoints such as the following:

- Number of days the patient is out of the intensive care unit.
- General quality of life.

The trial's sponsor is interested in developing a gatekeeping procedure that will enable us to test the statistical significance of the secondary endpoints after at least one primary outcome variable was found significant. This means that the primary family of hypotheses serves as a parallel gatekeeper for the secondary family.

**EXAMPLE: Case study 6 (Schizophrenia trial)**

This case study is based on a Phase III clinical trial for the treatment of schizophrenia. The trial was conducted to evaluate the efficacy of three doses of a new treatment (Doses L, M, and H) versus a placebo. Each dose-placebo comparison is performed with respect to the following three endpoints:

- Endpoint P (primary endpoint based on the mean change from baseline in the Positive and Negative Syndrome Scale total score at Week 6).
- Endpoint S1 (key secondary endpoint based on the mean change from baseline in the Clinical Global Impression-Severity score at Week 6).
- Endpoint S2 (key secondary endpoint based on the mean change from baseline in the Positive and Negative Syndrome Scale total score at Day 4).

The three endpoints are examined sequentially starting with Endpoint P and proceeding to Endpoints S1 and S2. If a dose is not shown to be superior to placebo for the primary endpoint, there is no value in further testing the efficacy of that dose for the two secondary endpoints. Similarly, if a dose is shown to be efficient for the primary endpoint but not for Endpoint S1, this dose will not be tested for Endpoint S2. Multiplicity problems of this kind with several sequences of hierarchical tests are known as problems with multiple branches.

### 5.6.2 General principles of gatekeeping inferences based on the mixture method

In order to explain how to set up gatekeeping procedures in the case studies presented above, we will describe the general framework for performing gatekeeping inferences based on the mixture method. (Technical details are provided in the Appendix to this chapter.)

Consider a clinical trial in which the sponsor is interested in testing  $n \geq 2$  null hypotheses  $H_1, \dots, H_n$  while controlling the FWER in the strong sense. Without loss of generality, we assume that the null hypotheses are grouped into two hierarchically ordered families denoted by  $F_1$  and  $F_2$ . These families will be referred to

as the primary and secondary families, respectively. The principle for constructing gatekeeping procedures with more than two families remains similar and will be illustrated later using the schizophrenia trial example from Case study 6.

### Logical restrictions

To reflect the hierarchical structure of the multiplicity problem, logical restrictions are defined among the null hypotheses in the primary and secondary families. These restrictions are applied to identify *testable* and *non-testable* hypotheses in the secondary family depending on the number of rejected hypotheses in the primary family. A hypothesis in Family  $F_1$  is termed testable if all hypotheses from Family  $F_1$  that it depends on were rejected. A non-testable hypothesis is accepted without testing even though the raw value associated with this hypothesis might be highly significant. As a quick illustration, in serial gatekeeping, hypotheses in Family  $F_2$  are only testable if all hypotheses are rejected in Family  $F_1$ . In parallel gatekeeping, hypotheses in Family  $F_2$  are testable if at least one hypothesis is rejected in Family  $F_1$ .

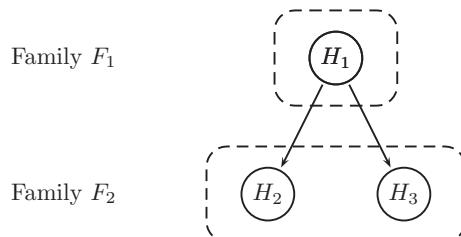
Using the depression trial from Case study 4 as an example, the three null hypotheses tested in this trial are defined as follows:

- $H_1$  (no treatment effect with respect to mean HAMD17 improvement)
- $H_2$  (no treatment effect with respect to HAMD17 response rate)
- $H_3$  (no treatment effect with respect to HAMD17 remission rate)

Also,  $p_1$ ,  $p_2$ , and  $p_3$  denote the one-sided  $p$ -values for testing the corresponding hypotheses. The FWER is set to  $\alpha = 0.025$  (one-sided) in this case study.

Two families of hypotheses are defined in the depression trial. Family  $F_1$  contains  $H_1$ , and Family  $F_2$  contains  $H_2$  and  $H_3$ . As shown in Figure 5.16, Family  $F_1$  serves as a serial gatekeeper for Family  $F_2$  because the hypotheses  $H_2$  and  $H_3$  can be tested only if  $H_1$  is rejected.

**Figure 5.16**  
Case study 4



*Two families of null hypotheses in the depression trial from Case study 4. Family  $F_1$  serves as a serial gatekeeper and the hypotheses in Family  $F_2$  will be tested only after the primary hypothesis was rejected.*

### Component procedures

The mixture approach will be used to set up a gatekeeping procedure that protects the overall FWER in the depression trial at the nominal level. In general, a mixture-based gatekeeping procedure is defined by specifying component procedures for each family of hypotheses. Any procedure that provides local error rate control can be considered, including the Bonferroni, Holm, Hochberg, Hommel, or other procedures. For example, the following component procedures will be used in Case study 4:

- Family  $F_1$ : Univariate test (no multiplicity adjustment is needed because there is only one null hypothesis in this family).
- Family  $F_2$ : Holm procedure.

Another important condition is that the component procedures need to be *separable*. A separable procedure can be thought of as a “generous” procedure that enables the process of transferring the error rate to the next family even if some null hypotheses are accepted in the current family. Note that the last family in the sequence does not need to be separable because the error rate does not need to be further transferred to any subsequent family. The Bonferroni procedure is known to be separable, but popular stepwise procedures, e.g., Holm, Hochberg and Hommel, are not separable. However, non-separable procedures can be modified to make them separable. These modified procedures are known as *truncated procedures* (Dmitrienko et al., 2008). The process of truncating a non-separable procedure will be illustrated using Case study 6.

### Intersection *p*-values

The mixture method is based on the closure principle (see Section 5.3). Consider an intersection hypothesis

$$H(I) = \bigcap_{i \in I} H_i$$

with  $I \subseteq \{1, \dots, n\}$  being an index set of all hypotheses in this multiplicity problem. Let  $I_1$  and  $I_2$  be the subsets of  $I$  containing the indexes belonging to Families  $F_1$  and  $F_2$ , respectively. Further, let  $I_2^*$  be the restricted index set defined from  $I_2$  by removing the indexes of hypotheses that are non-testable based on the logical restrictions.

The intersection *p*-value for  $H(I)$ , denoted by  $p(I)$ , is calculated using a mixture of the family-specific component procedures:

$$p(I) = \min \left[ \frac{p_1(I_1)}{c_1(I)}, \frac{p_2(I_2^*)}{c_2(I)} \right],$$

where  $p_1(I_1)$  and  $p_2(I_2^*)$  are the component *p*-values calculated using the family component procedures. Further,  $c_1(I)$  and  $c_2(I)$  define the fractions of the overall FWER applied to Families  $F_1$  and  $F_2$ , respectively. These quantities account for the hierarchical structure of this multiplicity problem. In general multiplicity problems with two families of hypotheses (primary and secondary hypotheses),  $c_1(I)$  is always set to 1, which means that the full  $\alpha = 0.025$  is used when the primary hypotheses are tested. However,  $c_2(I)$  may be less than 1. That is, the error rate used in Family  $F_2$  might be less than  $\alpha$ , depending on the outcomes of the primary tests.

Coming back to Case study 4, the process of calculating intersection *p*-values is illustrated below using two examples. Starting with the intersection hypothesis  $H_1 \cap H_2$ , we have  $I = \{1, 2\}$  with

$$I_1 = \{1\}, \quad I_2 = \{2\}.$$

When decision rules for intersection hypotheses are set up, a primary hypothesis included in an intersection is treated as if it failed to be rejected. In this particular case, since the intersection contains  $H_1$ , it follows from the serial logical restriction that the null hypothesis  $H_2$  is not testable and is automatically accepted. As a consequence, this secondary hypothesis is removed from the intersection, and the resulting restricted index set  $I_2^*$  is empty. The intersection *p*-value is then simply equal to the raw *p*-value for  $H_1$ , i.e.,

$$p(I) = p(\{12\}) = \frac{p_1(I_1)}{c_1(I)} = p_1$$

since, as stated above,  $c_1(I)$  is always equal to 1 and  $p_2(I_2^*)$  is excluded from the calculation.

Consider now the intersection hypothesis  $H_2 \cap H_3$ . It is easy to see that  $I = \{2, 3\}$ ,  $I_1$  is empty, and  $I_2 = \{2, 3\}$ . Since  $H_1$  is not included in this intersection, it is treated as a rejected hypothesis. Thus, based on the logical restriction, both  $H_2$  and  $H_3$  are testable and their indices are retained. The restricted index set is given by

$$I_2^* = \{2, 3\}.$$

Therefore, the intersection  $p$ -value is equal to

$$p(I) = p(\{23\}) = \frac{p_2(I_2^*)}{c_2(I)} = p_2(I_2^*),$$

where  $c_2(I) = 1$  since the serial gatekeeper (Family  $F_1$ ) is successfully passed and the full  $\alpha$  level is available for testing the hypotheses in Family  $F_2$ . Recalling that the component procedure in Family  $F_2$  is the Holm procedure, we have

$$p(I) = p(\{23\}) = 2 \min(p_2, p_3).$$

### Adjusted $p$ -values

After the intersection  $p$ -values were defined for all intersection hypotheses in the closed family, the multiplicity-adjusted  $p$ -value is computed for each null hypothesis. It follows from the closure principle that the adjusted  $p$ -value for the null hypothesis  $H_i$  is given by

$$\tilde{p}_i = \max p(I),$$

where the maximum is computed over all index sets  $I$  such that  $i \in I$ . The mixture gatekeeping procedure rejects  $H_i$  if and only if  $\tilde{p}_i \leq \alpha$ ,  $i = 1, \dots, n$ .

There are  $2^3 - 1 = 7$  intersection hypotheses in the closed family in Case study 4. The corresponding intersection  $p$ -values are given in Table 5.7. The adjusted  $p$ -value for the null hypothesis  $H_i$  is calculated by finding the largest intersection  $p$ -value over all index sets  $I$  such that  $i \in I$  in this table, i.e.,

$$\tilde{p}_1 = \max(p(\{123\}), p(\{12\}), p(\{13\}), p(\{1\})) = p_1,$$

$$\tilde{p}_2 = \max(p(\{123\}), p(\{12\}), p(\{23\}), p(\{2\})) = \max(p_1, 2 \min(p_2, p_3), p_2),$$

$$\tilde{p}_3 = \max(p(\{123\}), p(\{13\}), p(\{23\}), p(\{3\})) = \max(p_1, 2 \min(p_2, p_3), p_3).$$

This means that the hypothesis  $H_1$  is rejected by the gatekeeping procedure if  $\tilde{p}_1 = p_1 \leq \alpha$ . Similarly, the hypothesis  $H_2$  is rejected if  $\tilde{p}_2 \leq \alpha$  or, equivalently, if

$$p_1 \leq \alpha \text{ and } \min(p_2, p_3) \leq \alpha/2 \text{ and } p_2 \leq \alpha.$$

The resulting inferences guarantee strong control over the two families of hypotheses at a one-sided  $\alpha = 0.025$ .

**TABLE 5.7** Intersection  $p$ -values in Case study 4

Index set $I$	Restricted index set $I^*$	Intersection $p$ -value $p(I)$
{123}	{1}	$p_1$
{12}	{1}	$p_1$
{13}	{1}	$p_1$
{1}	{1}	$p_1$
{23}	{23}	$2 \min(p_2, p_3)$
{2}	{2}	$p_2$
{3}	{3}	$p_3$

### 5.6.3 Implementation of mixture-based gatekeeping procedures

In this section, we will introduce gatekeeping procedures constructed using the mixture method. The `%MixGate` macro will be used to illustrate the SAS implementation of the mixture method in Case studies 4 and 5.

#### `%MixGate` macro

The `%MixGate` macro implements commonly used  $p$ -value based gatekeeping procedures, computes adjusted  $p$ -values, and performs power computation. Gatekeeping procedures based on the following components are supported by this macro:

- Bonferroni
- Holm
- Hochberg
- Hommel

As mentioned previously, an important feature of gatekeeping procedures is that logical restrictions can be specified to account for clinically relevant relationships among the null hypotheses of interest. The `%MixGate` macro implements logical restrictions that are formulated in terms of serial and parallel rejection sets (Dmitrienko et al., 2007; Brechenmacher et al., 2011). The macro's arguments are listed below (with their possible values indicated in brackets and default value underlined):

- **Indata** [dataset name] specifies the name of the data set that defines the structure of the multiplicity problem, i.e., the individual null hypotheses, families and logical restrictions. More information on the required variables is provided below.
- **Method** [Standard, Modif] specifies whether the standard or modified mixture method is used to construct the gatekeeping procedure. The modified mixture method is introduced in Section 5.6.4.
- **Test** [Bonf, Holm, Hochberg, Hommel] defines the component procedure, i.e., the MTP applied within each family of hypotheses. If **Test**=Bonf, the Bonferroni procedure is used as the component procedure in each family except the last one, where the Holm procedure is applied. If **Test**=Holm, **Test**=Hochberg, or **Test**=Hommel, a truncated version of the component procedure is used in each family except the last family, where the regular procedure is used. A more powerful procedure is used in the last family since the condition of separability can be relaxed (see Section 5.6.2).
- **Gamma** [ $0 \leq \text{numerical} < 1$ , NULL] is the truncation parameter for each family (except the last family where the truncation parameter is automatically set to 1). Values are separated by #. For example, `gamma=0.5#0.9` can be used in a multiplicity problem with 3 families and indicates that  $\gamma_1 = 0.5$  and  $\gamma_2 = 0.9$  ( $\gamma_3$  is automatically set to 1). Note that this parameter does not need to be specified for Bonferroni-based gatekeeping procedures because the Bonferroni procedure is separable.
- **adjpout** [dataset name, AdjP] specifies the name of the data set created by the macro with the adjusted  $p$ -value for each hypothesis.
- **powout** [dataset name, Indpow] specifies the name of the data set created by the macro with marginal power values for each hypothesis (only used for power calculation).
- **Alpha** [ $0 < \text{numerical} < 1$ , 0.025] is the global FWER (only used for power calculations). In this setting, the FWER is typically one-sided.

- `rawp [dataset name, NULL]` specifies the name of the data set containing the raw *p*-values used to compute power (only used for power calculation). In this setting, the raw *p*-values are typically one-sided.
- `pvalout [dataset name, NULL]` specifies the name of the data set created by the macro with adjusted *p*-values (only used for power calculations).

Before calling the `%MixGate` macro, the testing strategy must be specified in a data set (which name will be indicated in the macro parameter `Indata`), i.e., the individual hypotheses, families and logical restrictions. This data set must contain one row per hypothesis and the following variables:

- `Hyp` is the label used for each individual hypothesis, e.g.,  $H_1$ .
- `Family` is the family index.
- `Parallel` is a set of Boolean indicators that define the indices of the null hypotheses included in the parallel set of the current hypothesis. The Boolean indicators are ordered based on the position of the corresponding hypotheses in the `Indata` data set. For example, consider a problem with three null hypotheses  $H_1$ ,  $H_2$ , and  $H_3$ , and assume that  $H_1$  and  $H_2$  are both in the parallel set of  $H_3$ . In this case, the `parallel` variable for  $H_3$  is set to 110.
- `Serial` is set up similarly to `parallel` and defines the indices of null hypotheses included in the serial set of the current hypothesis.
- `Rawp` is the raw *p*-value associated with each null hypothesis (not needed for power calculation).

## Bonferroni-based gatekeeping procedure in Case study 4

The depression trial example from Case study 4 was used earlier in this section to illustrate the general principles of gatekeeping inferences based on the mixture method. There are three null hypotheses in this trial. The first null hypothesis,  $H_1$ , serves as a serial gatekeeper for the other two hypotheses. Thus,  $H_2$  and  $H_3$  become testable only if  $H_1$  was rejected. As stated earlier, the mixture method is applied using the following component procedures:

- Family  $F_1$ : Univariate test.
- Family  $F_2$ : Holm procedure.

Program 5.14 demonstrates how to calculate the adjusted *p*-values for this gatekeeping procedure in Case study 4 using the `%MixGate` macro. The `study` data set, which will be referred to in the macro parameter `Indata`, contains three records, (one for each of the three null hypotheses), which are identified using the `hyp` variable. Since Family  $F_1$  includes one hypothesis and Family  $F_2$  includes two hypotheses: `family=1` in the first record and `family=2` in the second and third records. In this example, the primary family is a serial gatekeeper. Thus, there is no parallel logical restriction, and all indicators in the `parallel` variable are set to 0. The first indicator of the `serial` variable is set to 1 for the second and third records to indicate that  $H_2$  and  $H_3$  are both logically restricted by  $H_1$  in a serial manner. The `rawp` variable defines the two-sided *p*-values. The two-sided *p*-value for the primary analysis is  $p_1 = 0.046$  and the two-sided raw *p*-values associated with the HAMD17 response and remission rates are  $p_2 = 0.048$  and  $p_3 = 0.021$ , respectively.

**PROGRAM 5.14 Bonferroni-based gatekeeping procedure in Case study 4**

```

/* Testing strategy */
data study;
    input hyp $ family parallel $ serial $ rawp;
    datalines;
        H1 1 000 000 0.046
        H2 2 000 100 0.048
        H3 2 000 100 0.021
    run;

/* Bonferroni-based gatekeeping procedure */
%MixGate(Indata=study,method=Standard,test=Bonf,adjpout=adjp);

title "Bonferroni-based gatekeeping: Adjusted p-values";
proc print data=adjp noobs label;
    var adj_p1-adj_p3;
run;

```

**Output from Program 5.14**


---

Bonferroni-based gatekeeping: Adjusted p-values		
Adjusted p-value:	Adjusted p-value:	Adjusted p-value:
H1	H2	H3
0.0460	0.0480	0.0460

---

The adjusted  $p$ -values calculated by the macro are shown in Output 5.14. The three adjusted  $p$ -values are significant at the two-sided 5% level, which indicates that the experimental treatment demonstrates a significant effect on the primary and both secondary endpoints after a multiplicity adjustment based on this gatekeeping procedure.

It is interesting to note that in multiplicity problems with serial gatekeepers, the Bonferroni gatekeeping procedure has a simple stepwise version that leads to the same conclusions as above:

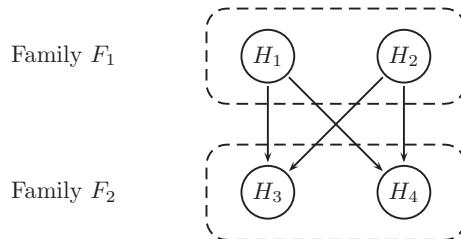
- Step 1. If  $p_1 \leq \alpha$ , reject  $H_1$  and go to Step 2. Otherwise, stop and accept all null hypotheses.
- Step 2.  $H_2$  and  $H_3$  are tested at the  $\alpha$  level using the Holm procedure. (See Section 5.3.3.)

**Bonferroni-based gatekeeping procedure in Case study 5**

The mixture method is used in this example to set up a gatekeeping strategy for the ARDS trial from Case study 5 involving two families of hypotheses with a parallel gatekeeper. Let  $H_1$  and  $H_2$  denote the null hypotheses of no treatment effect with respect to the number of ventilator-free days (VFD) and 28-day all-cause mortality. Also, denote the hypotheses associated with the secondary endpoints by  $H_3$  and  $H_4$ . Lastly, the associated two-sided raw  $p$ -values will be denoted by  $p_1$ ,  $p_2$ ,  $p_3$ , and  $p_4$ .

The gatekeeping strategy in Case study 5 is depicted in Figure 5.17. Family  $F_1$  contains the hypotheses  $H_1$  and  $H_2$ , and Family  $F_2$  includes  $H_3$  and  $H_4$ . Logical

restrictions are set up so that Family  $F_1$  serves as a parallel gatekeeper to Family  $F_2$ , i.e., the secondary hypotheses can be tested only if at least one primary hypothesis is rejected.



**Figure 5.17**  
**Case study 5**

Two families of null hypotheses in the ARDS trial from Case study 5. Family  $F_1$  is a parallel gatekeeper, and the secondary endpoints are analyzed only if at least one primary hypothesis is rejected.

The next step for constructing a mixture-based gatekeeping procedure is to define a component procedure for each of the family. As an illustration, the following component procedures will be used in Case study 5:

- Family  $F_1$ : Bonferroni procedure.
- Family  $F_2$ : Holm procedure.

Table 5.8 presents the intersection  $p$ -values calculated using the mixture method for the  $2^4 - 1 = 15$  intersection hypotheses in the closed family. The intersection hypothesis  $H_2 \cap H_3 \cap H_4$  will be used to illustrate the process of calculating the intersection  $p$ -values using the general principles of the mixture method described in Section 5.6.2. For this intersection hypothesis, we have

$$I = \{2, 3, 4\}, I_1 = \{2\}, I_2 = \{3, 4\}.$$

As indicated in Section 5.6.2, if a primary hypothesis is included in an intersection, this hypothesis is treated as if it was accepted. In this case,  $I_1 = \{2\}$ , which means that  $H_1$  is rejected and  $H_2$  is accepted in the primary family. Based on the parallel logical restriction, the hypotheses  $H_3$  and  $H_4$  are both testable since at least one primary hypothesis is rejected. This implies that the restricted index set in the secondary family is equal to the original index set, i.e.,  $I_2^* = I_2$ . Per the mixture

**TABLE 5.8** Intersection  $p$ -values in Case study 5

Index set $I$	Restricted index set $I^*$	Intersection $p$ -value $p(I)$
{1234}	{12}	$\min(p_1/0.5, p_2/0.5)$
{123}	{12}	$\min(p_1/0.5, p_2/0.5)$
{124}	{12}	$\min(p_1/0.5, p_2/0.5)$
{12}	{12}	$\min(p_1/0.5, p_2/0.5)$
{134}	{134}	$\min(p_1/0.5, p_3/0.25, p_4/0.25)$
{13}	{13}	$\min(p_1/0.5, p_3/0.5)$
{14}	{14}	$\min(p_1/0.5, p_4/0.5)$
{1}	{1}	$p_1/0.5$
{234}	{234}	$\min(p_2/0.5, p_3/0.25, p_4/0.25)$
{23}	{23}	$\min(p_2/0.5, p_3/0.5)$
{24}	{24}	$\min(p_2/0.5, p_4/0.5)$
{2}	{2}	$p_2/0.5$
{34}	{34}	$\min(p_3/0.5, p_4/0.5)$
{3}	{3}	$p_3$
{4}	{4}	$p_4$

method, the intersection  $p$ -value is then equal to:

$$p(I) = \min \left[ \frac{p_1(I_1)}{c_1(I)}, \frac{p_2(I_2^*)}{c_2(I)} \right].$$

As explained in Section 5.6.2,  $c_1(I)$  is always equal to 1, and  $c_2(I)$  corresponds to the remaining fraction of  $\alpha$  available after testing the hypotheses in Family  $F_1$ . Within this intersection hypothesis, only  $H_1$  is rejected, and its  $\alpha$  level ( $\alpha/2$  as per Bonferroni testing) can be carried over to Family  $F_2$ , which means that  $c_2(I) = 1/2$ . Given that  $p_1(I_1)$  and  $p_2(I_2^*)$  are the adjusted  $p$ -values obtained using the Bonferroni and Holm procedures applied to  $I_1$  and  $I_2^*$ , respectively, we have

$$p_1(I_1) = 2p_2 \text{ and } p_2(I_2^*) = 2 \min(p_3, p_4)$$

and, as a result,

$$p(I) = \min(p_2/0.5, p_3/0.25, p_4/0.25).$$

Program 5.15 uses the `%MixGate` macro to apply this gatekeeping procedure to perform a multiplicity adjustment in the ARDS trial under the following two scenarios:

- Scenario 1 assumes that the experimental treatment provides a statistically significant improvement with respect to both primary endpoints. The two-sided raw  $p$ -values are given by  $p_1 = 0.002$ ,  $p_2 = 0.024$ ,  $p_3 = 0.010$ , and  $p_4 = 0.005$ .
- Under Scenario 2, the analysis of the VFD endpoint yields a significant result, but the treatment difference with respect to 28-day all-cause mortality is only marginally significant. The two-sided raw  $p$ -values are given by  $p_1 = 0.002$ ,  $p_2 = 0.030$ ,  $p_3 = 0.010$ , and  $p_4 = 0.005$ .

The secondary analyses are assumed to produce highly significant results under both scenarios.

### PROGRAM 5.15 Bonferroni-based gatekeeping procedure in Case study 5

```
/* Testing strategy under Scenario 1 */
data study;
    input hyp $ family parallel $ serial $ rawp;
    datalines;
        H1 1 0000 0000 0.002
        H2 1 0000 0000 0.024
        H3 2 1100 0000 0.010
        H4 2 1100 0000 0.005
    run;

/* Bonferroni-based gatekeeping procedure */
%MixGate(Indata=study,method=Standard,test=Bonf,adjpout=adjp);

title "Bonferroni-based gatekeeping - Scenario 1: Adjusted p-values";
proc print data=adjp noobs label;
    var adj_p1-adj_p4;
run;

/* Testing strategy under Scenario 2 */
data study;
```

```

      input hyp $ family parallel $ serial $ rawp;
      datalines;
      H1 1 0000 0000 0.002
      H2 1 0000 0000 0.030
      H3 2 1100 0000 0.010
      H4 2 1100 0000 0.005
      run;

      /* Bonferroni-based gatekeeping procedure */
      %MixGate(Indata=study,method=Standard,test=Bonf,adjpout=adjp);

      title "Bonferroni-based gatekeeping - Scenario 2: Adjusted p-values";
      proc print data=adjp noobs label;
         var adj_p1-adj_p4;
      run;

```

---

**Output from Program 5.15** Bonferroni-based gatekeeping - Scenario 1: Adjusted p-values

Adjusted p-value: H1	Adjusted p-value: H2	Adjusted p-value: H3	Adjusted p-value: H4
0.0040	0.0480	0.0200	0.0200

Bonferroni-based gatekeeping - Scenario 2: Adjusted p-values

Adjusted p-value: H1	Adjusted p-value: H2	Adjusted p-value: H3	Adjusted p-value: H4
0.0040	0.0600	0.0200	0.0200

---

Output 5.15 lists the adjusted *p*-values under each treatment effect scenario. The adjusted *p*-values for the primary hypotheses are both significant at the two-sided 5% level in Scenario 1. Since the primary hypotheses are both rejected, the Bonferroni gatekeeping procedure proceeds to testing the secondary hypotheses, both of which are also rejected in this scenario. Scenario 2 presents a more interesting situation. The adjusted *p*-value associated with  $H_1$  is less than 0.05, whereas the adjusted *p*-value for  $H_2$  is greater than 0.05. This means that, in Scenario 2, only one of the primary comparisons is significant at the 5% level after a multiplicity adjustment (it is the comparison with respect to the VFD endpoint). Since the primary hypotheses are tested in a parallel manner, the Bonferroni gatekeeping procedure needs only one significant primary comparison to continue to the second stage and test the secondary hypotheses. The raw *p*-values associated with the secondary hypotheses are highly significant, and, as a consequence, the Bonferroni-based gatekeeping procedure successfully rejects both  $H_3$  and  $H_4$  in Scenario 2.

Similarly to the serial gatekeeping procedure from Case study 4, it is easy to show that the parallel gatekeeping used in Case study 5 also has a simple stepwise version, which enables the trial's sponsor to test the null hypotheses without having to examine all intersection hypotheses in the closed family:

- Step 1. Test  $H_1$  and  $H_2$  at  $\alpha = 0.05$  using the Bonferroni procedure. If at least one hypothesis is rejected, go to Step 2. Otherwise, accept all hypotheses and stop.

- Step 2. Let  $k$  be the number of hypotheses rejected in Step 1 ( $k = 1, 2$ ). Test  $H_3$  and  $H_4$  at  $k\alpha/2$  using the Holm procedure.

### Hommel-based gatekeeping procedure in Case study 6

In Case studies 4 and 5, basic Bonferroni-based gatekeeping procedures were employed and logical restrictions were defined using only one serial or parallel gatekeeper. The schizophrenia trial from Case study 6 presents a more complex setting, which will be used to demonstrate the process of constructing more powerful gatekeeping procedures derived from the Hommel procedure. In this example, there are three doses (Doses L, M, and H) of an experimental treatment to be tested versus a placebo with respect to the primary endpoint (Endpoint P) and two key secondary endpoints (Endpoints S1 and S2). The resulting nine null hypotheses of no effect are denoted by  $H_1, \dots, H_9$ . The null hypotheses are grouped into three families corresponding to the three endpoints, and the families are tested sequentially beginning with the first one:

$$F_1 = \{H_1, H_2, H_3\} \text{ (Endpoint P),}$$

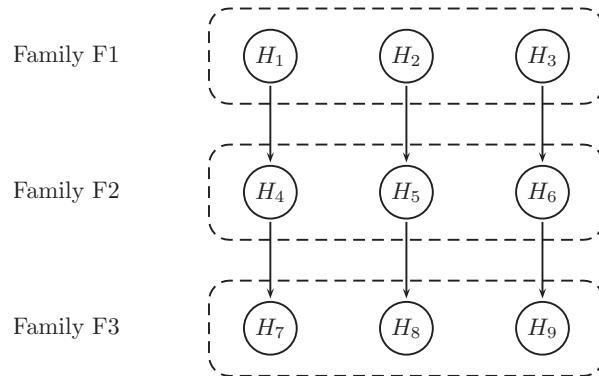
$$F_2 = \{H_4, H_5, H_6\} \text{ (Endpoint S1),}$$

$$F_3 = \{H_7, H_8, H_9\} \text{ (Endpoint S2).}$$

Here the null hypotheses  $H_1$ ,  $H_4$ , and  $H_7$  correspond to the comparisons of Dose L versus placebo;  $H_2$ ,  $H_5$ , and  $H_8$  correspond to the comparisons of Dose M versus placebo, and, lastly,  $H_3$ ,  $H_6$ , and  $H_9$  correspond to the comparisons of Dose H versus placebo.

As shown in Figure 5.18, the logical restrictions in this trial are defined using three separate sequences. A null hypothesis is testable only in the event that all preceding hypotheses in the sequence are rejected and non-testable hypotheses are automatically accepted without testing. For example, if the hypothesis  $H_1$  is not rejected, there is no evidence of a significant treatment effect for the primary endpoint at Dose L, and it is no longer clinically relevant to evaluate the treatment effect for the secondary endpoints at the same dose. This means that the other two hypotheses in this sequence ( $H_4$  and  $H_7$ ) are not testable.

**Figure 5.18**  
Case study 6



*Three families of hypotheses in the schizophrenia trial from Case study 6. The nine hypotheses are organized into three testing sequences corresponding to each dose-placebo comparison. Serial logical restrictions are imposed within each sequence to account for the clinically meaningful relationships.*

The component procedures for the three families are based on the Hommel procedure as it provides a power advantage compared to other  $p$ -value based

procedures such as the Bonferroni, Holm, or Hochberg procedures (see Section 5.3). In this clinical trial example, the Hommel procedure protects the error rate within each family because the hypothesis test statistics in each family follow a trivariate normal distribution with positive correlations.

As stated earlier, an important condition associated with the mixture method is that component procedures used in Families  $F_1$  and  $F_2$  must be separable. As the Hommel procedure is not separable, a truncated version of this procedure is defined by deriving a convex combination of its critical values with those of the Bonferroni procedure. (See the Appendix for the definition of the truncated Hommel procedure.) The truncated Hommel procedure is separable if the truncation parameter, denoted by  $\gamma$ , is less than 1. A truncated Hommel procedure with a prespecified truncation parameter  $\gamma$  will be denoted by  $\text{Hommel}(\gamma)$ . Note that  $\text{Hommel}(1)$  is equivalent to the regular Hommel procedure.

The Hommel-based gatekeeping procedure in Case study 6 uses the following component procedures with the prospectively defined truncation parameters  $\gamma_1$  and  $\gamma_2$  ( $0 \leq \gamma_1 < 1$  and  $0 \leq \gamma_2 < 1$ ) in the three families:

- Family  $F_1$ : Truncated Hommel procedure:  $\text{Hommel}(\gamma_1)$ .
- Family  $F_2$ : Truncated Hommel procedure:  $\text{Hommel}(\gamma_2)$ .
- Family  $F_3$ : Regular Hommel procedure:  $\text{Hommel}(1)$ .

To define the decision rules in the individual hypotheses in this multiplicity problem, the closed family consisting of all non-empty intersections of  $H_1, \dots, H_9$  is set up first. There is no simple stepwise algorithm available, and all  $2^9 - 1 = 511$  intersection  $p$ -values need to be calculated, which highlights the importance of efficient software implementation for the Hommel-based gatekeeping procedure.

The gatekeeping procedure is constructed following the general principles of the mixture method. Let  $H_I$ ,  $I \subseteq \{1, \dots, 9\}$  denote an arbitrary intersection hypothesis from the closed family. Using the notation introduced in Section 5.6.2, the formula for calculating the  $p$ -value associated with  $H_I$  is easily extended from the case of two families to three families as follows:

$$p(I) = \min \left[ \frac{p_1(I_1)}{c_1(I)}, \frac{p_2(I_2^*)}{c_2(I)}, \frac{p_3(I_3^*)}{c_3(I)} \right].$$

The derivations of the component  $p$ -values (i.e.,  $p_1(I_1)$ ,  $p_2(I_2^*)$  and  $p_3(I_3^*)$ ), and  $c_i(I)$ ,  $i = 1, 2, 3$ , used in the Hommel-based gatekeeping procedure are given in the Appendix.

To illustrate the Hommel-based gatekeeping procedure using the `%MixGate` macro, the truncation parameters will be set to  $\gamma_1 = 0.5$  and  $\gamma_2 = 0.5$  for simplicity. Program 5.16 shows how to set up the testing strategy in this example with multiple sequences of hypotheses, and it compares the adjusted  $p$ -values calculated using the Hommel-based and Bonferroni-based gatekeeping procedures for the following set of one-sided raw  $p$ -values:

- Family  $F_1$  (Endpoint P):  $p_1 = 0.1403$ ,  $p_2 = 0.0095$ ,  $p_3 = 0.0088$ .
- Family  $F_2$  (Endpoint S1):  $p_4 = 0.0559$ ,  $p_5 = 0.0051$ ,  $p_6 = 0.0021$ .
- Family  $F_3$  (Endpoint S2):  $p_7 = 0.1781$ ,  $p_8 = 0.0015$ ,  $p_9 = 0.0008$ .

This configuration of  $p$ -values corresponds to a case where there is no evidence of efficacy at the low dose ( $p$ -values are non-significant across the three endpoints), but the treatment provides a beneficial effect at the medium and high doses (most  $p$ -values are highly significant).

**PROGRAM 5.16 Hommel-based gatekeeping procedure in Case study 6**

```

/* Testing strategy */
data study;
    input hyp $ family parallel $ serial $ rawp;
    datalines;
    H1 1 000000000 000000000 0.1403
    H2 1 000000000 000000000 0.0095
    H3 1 000000000 000000000 0.0088
    H4 2 000000000 100000000 0.0559
    H5 2 000000000 010000000 0.0051
    H6 2 000000000 001000000 0.0021
    H7 3 000000000 100100000 0.1781
    H8 3 000000000 010010000 0.0015
    H9 3 000000000 001001000 0.0008
run;

/*Bonferroni-based gatekeeping procedure */
%MixGate(Indata=study,method=Standard,test=Bonf,
adjpout=adjp1);

/*Hommel-based gatekeeping procedure */
%MixGate(Indata=study,method=Standard,test=Hommel,
gamma=0.5#0.5,adjpout=adjp2);

proc transpose data=adjp1 out=t_adjp1 prefix=bonf;
    var adj_p1-adj_p9;
run;

proc transpose data=adjp2 out=t_adjp2 prefix=hommel;
    var adj_p1-adj_p9;
run;

data adjp (drop=_NAME_);
    merge t_adjp1 t_adjp2;
    by _NAME_;
    label _LABEL_="Null hypotheses" bonf1="Bonferroni-based"
        hommeli="Hommel-based";
run;

title "Bonferroni and Hommel-based gatekeeping: Adjusted p-values";
proc print data=adjp noobs label;
run;

```

**Output from Program 5.16**

Bonferroni and Hommel-based gatekeeping: Adjusted p-values

Null hypotheses	Bonferroni-based	Hommel-based
Adjusted p-value: H1	0.4209	0.2105
Adjusted p-value: H2	0.0285	0.0228
Adjusted p-value: H3	0.0264	0.0211
Adjusted p-value: H4	0.4209	0.2105
Adjusted p-value: H5	0.0285	0.0230

Adjusted p-value: H6	0.0264	0.0211
Adjusted p-value: H7	0.4209	0.2105
Adjusted p-value: H8	0.0285	0.0230
Adjusted p-value: H9	0.0264	0.0228

Output 5.16 presents the adjusted  $p$ -values obtained using the Bonferroni-based and Hommel-based gatekeeping procedures. Individual hypothesis tests are performed at a one-sided  $\alpha = 0.025$ . It can be seen that all adjusted  $p$ -values calculated using the Hommel-based gatekeeping procedure are smaller in magnitude compared to those obtained using the Bonferroni-based method. As a result, the Hommel-based gatekeeping procedure successfully rejects a large number of hypotheses and establishes a significant treatment effect at the medium and high doses for all three endpoints (the hypotheses  $H_2$ ,  $H_3$ ,  $H_5$ ,  $H_6$ ,  $H_8$ , and  $H_9$  are all rejected). The Bonferroni-based gatekeeping procedure rejects no hypotheses since all adjusted  $p$ -values are greater than 0.025. In this example, the Hommel-based gatekeeping procedure provides a clear advantage over the Bonferroni-based method.

The truncation parameters  $\gamma_1$  and  $\gamma_2$  were set to 0.5 to help illustrate the implementation of the gatekeeping procedure in Program 5.16. It is important to note that these parameters help control the balance between the power levels in the three families of hypotheses. Taking  $\gamma_1$  as an example, when this parameter is set to a value close to 1, the procedure used in Family  $F_1$  is very close to the regular Hommel procedure, which increases power of the primary endpoint tests but, at the same time, might decrease power of the secondary endpoint tests. On the other hand, with  $\gamma_1 = 0$ , the truncated Hommel procedure simplifies to the Bonferroni procedure, which leads to lower power in Family  $F_1$  but might improve power in subsequent families. Optimal selection of the truncation parameters is needed to ensure a trade-off among the probabilities of success in the three families and can be achieved through clinical trial simulations at the trial planning stage (Brechenmacher et al., 2011; Dmitrienko et al., 2011; Dmitrienko et al., 2015). For more information on general approaches to clinical trial optimization with applications to optimal multiplicity adjustments in Phase III trials with complex objectives, see Dmitrienko and Pulkstenis (2017).

#### 5.6.4 Modified mixture method

A new formulation of the mixture method for constructing gatekeeping procedures designed for multiplicity problems where the null hypotheses of interest are arranged into several branches was developed in Dmitrienko et al. (2016). This class of multiplicity problems was defined in Dmitrienko et al. (2009) and applies to very common situations in Phase III programs with ordered endpoints, including, for example, Case study 6. It was shown in Dmitrienko et al. (2016) that the modified mixture method leads to a uniform power gain compared to the standard mixture method for this class of multiplicity problems.

The modified method is conceptually similar to the standard method in that it also relies on the closure principle. The difference lies in the rules for defining intersection  $p$ -values in the closed family. In fact, the trial-specific logical restrictions play a more prominent role in the modified method compared to the standard method. The mathematical details and key differences between the standard and modified mixture methods are given in the Appendix to this chapter. The `%MixGate` macro can be used for implementing the standard method as well as the modified method by setting `method=Standard` and `method=Modif`, respectively.

In this subsection, the `%MixGate` macro will be used to quantify the power gain of the modified method versus the standard method when applied to the Hommel-based

gatekeeping procedure in the setting of the schizophrenia trial from Case study 6. The process of power calculation in clinical trials with multiple objectives follows the same general principles used in trials with a single primary objective. General considerations for sample size and power calculations in the setting of multiple null hypotheses have been discussed in the literature. See, for example, Senn and Bretz (2007), Bretz et al. (2011b), Dmitrienko et al. (2011), Dmitrienko and D'Agostino (2013), and Dmitrienko et al. (2015). It is important to note that in clinical trials with multiple objectives, there are in general no direct formulas available to compute power, and it is usually calculated using simulation-based methods.

Using the schizophrenia trial example from Case study 6, simulations will be performed to assess the power gain of the modified mixture method over the standard method using the following assumptions. A multivariate normal distribution for the test statistics and a balanced design with  $n = 120$  patients per treatment arm is assumed. Further, the treatment effect is assumed to be present across all endpoints at Doses M and H only. The effect sizes for the corresponding tests are set to 0.37, which guarantees 80% marginal power for each test at a one-sided  $\alpha = 0.025$  (the effect size is defined as the mean treatment difference divided by the common standard deviation). The following values of the pairwise correlations among the three endpoints will be assumed:

$$\rho(P, S1) = 0.8, \rho(P, S2) = 0.4, \rho(S1, S2) = 0.3.$$

Based on these assumptions, Program 5.17 calls the `%MultiSeqSimul` macro to create a set of 10,000 simulated one-sided raw  $p$ -values that will be used in power calculations. This macro creates data sets that can be used in clinical trial simulations. It requires the following arguments:

- `n_htest`: Number of tests per family.
- `n_fam`: Number of families.
- `n_subj`: Number of patients per trial arm (assuming a balanced design).
- `out`: The name of the output data set with simulated raw  $p$ -values.
- `n_sim`: Number of simulation runs.
- `seed`: Seed to be used in the simulations.
- `eff_size`: Vector of effect sizes (families are separated by #).
- `sigma`: Correlation matrix (families are separated by #)

This macro is used in Program 5.17 in the context of Case study 6 with three branches and three families of hypotheses, but, in general, the `%MultiSeqSimul` macro can be used with any number of sequences or families of null hypotheses. The `rawp` data set created by this program contains 10,000 rows, with each row corresponding to one simulation run, and nine variables each corresponding to the nine hypotheses.

### PROGRAM 5.17 Simulation data set in Case study 6

```
%MultiSeqSimul(n_Htest=3,n_fam=3,n_subj=120,n_sim=10000,
               out=rawp,seed=1234,
               eff_size=0 0.37 0.37#0 0.37 0.37#0 0.37 0.37,
               sigma=1 0.8 0.4#0.8 1 0.3#0.4 0.3 1);
```

The `%MixGate` macro can now be called to calculate marginal power for each individual null hypothesis using the `rawp` data set. Note that this macro creates a data set, which name is specified in the macro argument `pvalout`, that contains

the adjusted  $p$ -values based on the specified gatekeeping procedure for the 10,000 simulations. This enables the user to calculate power based on any custom success criterion.

Program 5.18 invokes the `%MixGate` macro to compare operating characteristics of the Hommel-based gatekeeping procedure with  $\gamma_1 = \gamma_2 = 0.5$  using the standard and modified methods. The characteristics of interest include marginal power of each individual null hypothesis as well as the following custom success criterion at a one-sided global FWER of  $\alpha = 0.025$ :

- Success criterion: Probability to reject at least one hypothesis in Family  $F_1$ , at least one hypothesis in Family  $F_2$ , and at least one hypothesis in Family  $F_3$ .

The custom success criterion is evaluated using the `%Custom` macro defined within Program 5.18. This success criterion reflects a typical set of goals in clinical trials with several endpoints. See Brechenmacher et al. (2011). Other popular approaches to defining success criteria in trials with complex clinical objectives are discussed in Dmitrienko and Pulkstenis (2017).

### **PROGRAM 5.18 Power comparison of the standard and modified methods**

```
/* Testing strategy */
data study;
    input hyp $ family parallel $ serial $;
    datalines;
    H1 1 000000000 000000000
    H2 1 000000000 000000000
    H3 1 000000000 000000000
    H4 2 000000000 100000000
    H5 2 000000000 010000000
    H6 2 000000000 001000000
    H7 3 000000000 100100000
    H8 3 000000000 010010000
    H9 3 000000000 001001000
run;

/*Hommel-based gatekeeping procedure */
/* Marginal power of the 9 null hypothesis tests */
%MixGate(Indata=study,method=Standard,test=Hommel,
          gamma=0.5#0.5,rawp=rawp,powout=indpow1,
          pvalout=adjp1);
%MixGate(Indata=study,method=Modif,test=Hommel,
          gamma=0.5#0.5,rawp=rawp,powout=indpow2,
          pvalout=adjp2);

proc transpose data=indpow1 out=t_pow1 prefix=standard;
    var hyp1-hyp9;
run;

proc transpose data=indpow2 out=t_pow2 prefix=modified;
    var hyp1-hyp9;
run;

data indpow (drop=_NAME_);
    merge t_pow1 t_pow2;
    by _name_;

```

```

label _label_="Null hypotheses" standard1="Standard method"
modified1="Modified method";
run;

title "Standard and modified mixture methods: Marginal power (%)";
proc print data=indpow noobs label;
run;

/* Power for the custom success criterion */
%macro Custom(method=,label=);
  data critpow&method (keep=_label_ critpow&method);
    set adjp&method;
    _label_="Success criterion";
    crit=(((adj_p1<=0.025)+(adj_p2<=0.025) +
           (adj_p3<=0.025))>=1)
      *(((adj_p4<=0.025)+(adj_p5<=0.025) +
         (adj_p6<=0.025))>=1)
      *(((adj_p7<=0.025)+(adj_p8<=0.025) +
         (adj_p9<=0.025))>=1);
    if crit=1 then critpow&method+1;
    if simnum=10000;
    critpow&method=(critpow&method*100)/10000;
    label critpow&method="&label method" _label_="Power";
  run;
%mend;

%Custom(method=1,label=Standard);
%Custom(method=2,label=Modified);

data critpow;
  merge critpow1 critpow2;
  by _label_;
run;

title "Standard and modified mixture methods:
Power (%) for the custom success criterion";
proc print data=critpow noobs label;
run;

```

**Output from  
Program 5.18**

Null hypotheses	Standard method	Modified method
Individual power: H1	1.60	1.60
Individual power: H2	70.58	70.58
Individual power: H3	69.67	69.67
Individual power: H4	0.64	0.64
Individual power: H5	54.50	56.52
Individual power: H6	53.84	55.76
Individual power: H7	0.14	0.14
Individual power: H8	32.15	40.54
Individual power: H9	31.21	40.26

Standard and modified mixture methods: Power (%) for the custom success criterion		
Power	Standard method	Modified method
Success criterion	43.29	52.9

Output 5.18 presents the results of the power calculations for the Hommel-based gatekeeping procedure based on the standard and modified mixture methods. It can be seen from the output that the modified mixture method ensures a uniform power advantage compared to the standard method both in terms of the rejection probabilities for the individual hypotheses as well as the custom success criterion. (Note that the modified and standard methods provide the same power for the null hypotheses  $H_1$ ,  $H_2$ , and  $H_3$  since, as shown in the Appendix, the two methods rely on the same set of tests in the first family.) In particular, looking at the success criterion and the null hypotheses in Family  $F_3$ , the modified mixture method is much more efficient than the standard method with the absolute power gain approaching 10%. In this example, the truncation parameters  $\gamma_1$  and  $\gamma_2$  were set to 0.5. However, as demonstrated in Dmitrienko et al. (2016), the magnitude of power advantage of the modified mixture method is not affected by the truncation parameters.

### 5.6.5 Summary

In this section, we described methods for addressing multiplicity issues arising in clinical trials with multiple families of hypotheses that are ordered in a hierarchical manner. Efficient solutions for multiplicity problems of this kind can be obtained using gatekeeping strategies based on the mixture method. We first reviewed the general framework for performing gatekeeping inferences and then illustrated it using examples from clinical trials with ordered endpoints and the following types of logical restrictions:

- A gatekeeping procedure with serial logical restrictions passes a gatekeeper only after it rejected all of the individual hypotheses within this gatekeeper. It is worth noting that fixed-sequence procedures described in Section 5.4 serve as an example of a serial gatekeeping procedure. Specifically, a serial gatekeeping procedure simplifies to the fixed-sequence procedure when each gatekeeper contains only one hypothesis.
- A gatekeeping procedure with parallel logical restrictions passes a gatekeeper if at least one of the hypotheses within the gatekeeper is rejected.
- A gatekeeping procedure with general logical restrictions was illustrated using a multiplicity problem commonly encountered in clinical trials where null hypotheses are arranged in multiple sequences of hypotheses. In this setting, serial logical restrictions are imposed within each sequence, and a null hypothesis will be tested only if all preceding null hypotheses in the sequence were rejected.

We also discussed that the mixture method can be used for constructing gatekeeping procedures based on simple single-step procedures (e.g., Bonferroni) as well as more powerful MTPs such as the Hommel procedure. The following gatekeeping procedures were introduced in this section and were implemented using the `%MixGate` macro:

- A Bonferroni-based gatekeeping procedure is a good choice when there is no information available on the joint distribution of the hypothesis test statistics. Otherwise, alternative gatekeeping procedures based on more powerful MTPs exist and should be investigated. In particular, the Hommel-based gatekeeping procedure was introduced in a complex clinical trial setting and was shown to provide power advantage over the Bonferroni-based gatekeeping procedure.
- A modified mixture method was also introduced as an efficient tool for setting up gatekeeping procedures in multiplicity problems where the null hypotheses are organized in several sequences of hypotheses. In this setting, the modified mixture method is uniformly more powerful than the standard mixture method. The magnitude of power gain was evaluated via simulations using the %MixGate and %MultiSeqSimul macros.

## 5.7 References

---

- Alosh, M., Huque, M.F. (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* 28, 3-23.
- Alosh, M., Bretz, F., Huque, M.F. (2014). Advanced multiplicity adjustment methods in clinical trials. *Statistics in Medicine* 33, 693-713.
- Bauer, P., Röhmel, J., Maurer, W., Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* 17, 2133-2146.
- Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A new and powerful approach to multiple testing. *Journal of Royal Statistical Society Series B* 57, 1289-1300.
- Benjamini, Y., Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics* 24, 407-418.
- Brechenmacher, T., Xu, J., Dmitrienko, A., Tamhane, A.C. (2011). A mixture gatekeeping procedure based on the Hommel test for clinical trial applications. *Journal of Biopharmaceutical Statistics* 21, 748-767.
- Bretz, F., Maurer, W., Brannath, W., Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 28, 586-604.
- Bretz, F., Posch, M., Glimm, E., Klingmueller, F., Maurer, W., Rohmeyer, K. (2011a). Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal* 53(6), 894-913.
- Bretz, F., Maurer, W., Hommel, G. (2011b). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine* 30(13), 1489-1501.
- Dmitrienko, A., Offen, W., Westfall, P.H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* 22, 2387-2400.
- Dmitrienko, A., Wiens, B.L. Tamhane, A.C., Wang, X. (2007). Tree-structured-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine* 26, 2465-2478.
- Dmitrienko, A., Tamhane, A.C., Wiens, B.L. (2008). General multistage gatekeeping procedures. *Biometrical Journal* 50, 667-677.
- Dmitrienko, A., Bretz, F., Westfall, P.H., Troendle, J., Wiens, B.L., Tamhane, A.C., Hsu, J.C. (2009). Multiple testing methodology. *Multiple Testing Problems in Pharmaceutical Statistics*. Dmitrienko, A., Tamhane, A.C., Bretz, F. (editors). New York: Chapman and Hall/CRC Press.

- Dmitrienko, A., Tamhane, A.C. (2011). Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. *Statistics in Medicine* 31(13), 1473-1488.
- Dmitrienko, A., Millen B., Brechenmacher T., Paux G. (2011). Development of gatekeeping strategies in confirmatory clinical trials. *Biometrical Journal* 53, 875-893.
- Dmitrienko, A., D'Agostino, R.B. (2013). Tutorial in Biostatistics: Traditional multiplicity adjustment methods in clinical trials. *Statistics in Medicine* 32, 5172-5218.
- Dmitrienko, A., D'Agostino, R.B., Huque, M.F. (2013). Key multiplicity issues in clinical drug development. *Statistics in Medicine* 32, 1079-1111.
- Dmitrienko, A., Tamhane, A.C. (2013). General theory of mixture procedures for gatekeeping. *Biometrical Journal* 55, 402-419.
- Dmitrienko, A., Paux, G., Brechenmacher, T. (2015). Power calculations in clinical trials with complex clinical objectives. *Journal of the Japanese Society of Computational Statistics* 28, 15-50.
- Dmitrienko, A., Kordzakhia, G., Brechenmacher, T. (2016). Mixture-based gatekeeping procedures for multiplicity problems with multiple sequences of hypotheses. *Journal of Biopharmaceutical Statistics* 26, 758-780.
- Dmitrienko, A., Pulkstenis, E. (2017) (editors). *Clinical Trial Optimization Using R*. New York: Chapman and Hall/CRC Press.
- Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50, 1096-1121.
- Dunnett, C.W., Tamhane, A.C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine* 10, 939-947.
- Dunnett, C.W., Tamhane, A.C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association* 87, 162-170.
- EMA. (2002). Points to consider on multiplicity issues in clinical trials. European Medicines Agency/Committee for Medicinal Products for Human Use.
- EMA. (2017). Guideline on multiplicity issues in clinical trials (draft). European Medicines Agency/Committee for Human Medicinal Products.
- FDA (2017). Draft guidance for industry: Multiple endpoints in clinical trials. U.S. Food and Drug Administration.
- Finner, H., Strassburger, K. (2002). The partitioning principle: a powerful tool in multiple decision theory. *The Annals of Statistics* 30, 1194-1213.
- Goeman, J.J., Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics* 38, 3782-3810.
- Hochberg, Y., Tamhane, A.C. (1987). *Multiple Comparison Procedures*. New York: John Wiley.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika* 75, 800-802.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65-70.
- Hommel, G. (1983). Test of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal* 25, 423-430.
- Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika* 33, 321-336.
- Hommel, G. (1988). A stagewise rejective multiple procedure based on a modified Bonferroni test. *Biometrika* 75, 383-386.

- Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika* 76, 624-625.
- Holland, B.S., Copenhaver, M.D. (1987). An improved sequentially rejective Bonferroni test procedure. *Biometrics* 43, 417-423. Correction in *Biometrics* 43, 737.
- Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*. London, UK: Chapman and Hall.
- Hsu, J.C., Berger, R.L. (1999). Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *Journal of the American Statistical Association* 94, 468-482.
- Liu, W. (1996). Multiple tests of a non-hierarchical finite family of hypotheses. *Journal of Royal Statistical Society Series B* 58, 455-461.
- Liu, Y., Hsu, J.C. (2009). Testing for efficacy in primary and secondary endpoints by partitioning decision paths. *Journal of the American Statistical Association* 104, 1661-1670.
- Marcus, R., Peritz, E., Gabriel, K.R. (1976). On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* 63, 655-660.
- Maurer, W., Hothorn, L., Lehmacuer, E. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a priori ordered hypotheses. *Biometrie in der Chemisch-in-Pharmazeutischen Industrie*. Vollman, J. (editor). Fischer-Verlag: Stuttgart.
- Millen, B., Dmitrienko, A. (2011). Chain procedures: a class of flexible closed testing procedures with clinical trial applications. *Statistics in Biopharmaceutical Research* 3, 14-30.
- Naik, U.D. (1975). Some selection rules for comparing p processes with a standard. *Communications in Statistics. Series A* 4, 519-535.
- Rüger, B. (1978). Das maximale Signifikanzniveau des Tests "Lehne  $H_0$  ab, wenn  $k$  unter  $n$  gegebenen Tests zur Ablehnungsführen". *Metrika* 25, 171-178.
- Sarkar, S.K., Chang, C.K. (1997). Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 92, 1601-1608.
- Sarkar, S.K. (1998) Some probability inequalities for censored MTP2 random variables: a proof of the Simes conjecture. *The Annals of Statistics* 26, 494-504.
- Sarkar, S.K. (2008). On the Simes inequality and its generalization. *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*. Balakrishnan, N., Pena, E.A., Silvapulle, M.J. (editors). Beachwood, OH: Institute of Mathematical Statistics.
- Senn, S., Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 6, 161-170.
- Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 62, 626-633.
- Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 63, 655-660.
- Stefansson, G., Kim, W.C., Hsu, J.C. (1988). On confidence sets in multiple comparisons. *Statistical Decision Theory and Related Topics IV*. Gupta, S.S., Berger, J.O. (editors). New York: Academic Press.
- Tamhane, A.C., Liu, L. (2008). On weighted Hochberg procedures. *Biometrika* 95, 279-294.
- Westfall, P.H., Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. New York: John Wiley.

- Westfall, P.H., Krishen, A. (2001). Optimally weighted, fixed sequence, and gate-keeping multiple testing procedures. *Journal of Statistical Planning and Inference* 99, 25-40.
- Westfall, P.H., Bretz, F. (2010). Multiplicity in Clinical Trials. *Encyclopedia of Biopharmaceutical Statistics*. Third Edition. Chow, S.C. (editor). New York: Marcel Decker Inc.
- Westfall, P.H., Tobias, R.D., Wolfinger, R.D. (2011). *Multiple Comparisons and Multiple Tests Using SAS*. Second Edition. Cary, NC: SAS Institute, Inc.
- Wiens, B.L. (2003). A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* 2, 211-215.
- Wiens, B.L., Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics* 15, 929-942.

## 5.8 Appendix

---

### Chain procedures: general algorithm

---

Consider the problem of testing the null hypotheses  $H_1, \dots, H_m$  using a chain procedure. The FWER must be protected in the strong sense at a predefined  $\alpha$ , e.g., a one-sided  $\alpha = 0.025$ . The general chain procedure is based on a serial algorithm (Bretz et al., 2009) that relies on a predefined ordering of hypotheses with possibility of retesting null hypotheses. To set up the testing algorithm, let  $w_i$  be the weights assigned to hypotheses  $H_i$ , and let  $\alpha_i = w_i\alpha$  denote the initial significance level for the hypothesis  $H_i$ ,  $i = 1, \dots, m$  where

$$\begin{aligned} w_i &\geq 0, \quad i = 1, \dots, m, \\ \sum_{i=1}^m w_i &= 1. \end{aligned}$$

The initial matrix of transition parameters that define the process of  $\alpha$  propagation is given by

$$\begin{pmatrix} g_{11} & \dots & g_{1m} \\ \dots & \dots & \dots \\ g_{m1} & \dots & g_{mm} \end{pmatrix}.$$

The transition parameters define the rule for updating the significance levels of the non-rejected hypotheses after each rejection and are assumed to satisfy the following condition:

$$g_{ii} = 0, \quad \sum_{j=1}^m g_{ij} = 1, \quad i = 1, \dots, m.$$

Finally, let  $M$  denote the index set of non-rejected hypotheses. Note that  $M = \{1, \dots, m\}$  at the beginning of the algorithm. The serial testing algorithm is defined as follows:

- Step 1. Reject  $H_1$  if  $p_1 \leq \alpha_1$ . If this hypothesis is rejected, remove its index from the index set  $M$  and update the significance levels and transition parameters for the non-rejected hypotheses as follows

$$\alpha_j = \alpha_j + \alpha_1 g_{1j}, \quad j \in M,$$

$$g_{jk} = \frac{g_{jk} + g_{j1}g_{1k}}{1 - g_{j1}g_{1j}}, j, k \in M; g_{jk} = 0 \text{ otherwise.}$$

If there are any non-rejected hypotheses, proceed to Step 2; otherwise, testing stops.

- Step  $i = 2, \dots, m - 1$ . Reject  $H_i$  if  $p_i \leq \alpha_i$ . If this hypothesis is rejected, remove its index from the index set  $M$  and update the significance levels and transition parameters for the non-rejected hypotheses as follows

$$\alpha_j = \alpha_j + \alpha_i g_{ij}, j \in M,$$

$$g_{jk} = \frac{g_{jk} + g_{ji}g_{ik}}{1 - g_{ji}g_{ij}}, j, k \in M; g_{jk} = 0 \text{ otherwise.}$$

If there are any non-rejected hypotheses, proceed to Step  $i + 1$ ; otherwise, testing stops.

- Step  $m$ . Reject  $H_m$  if  $p_m \leq \alpha_m$ . If this hypothesis is rejected, remove its index from the index set  $M$  and update the significance levels and transition parameters for the non-rejected hypotheses as follows

$$\alpha_j = \alpha_j + \alpha_m g_{mj}, j \in M,$$

$$g_{jk} = \frac{g_{jk} + g_{jm}g_{mk}}{1 - g_{jm}g_{mj}}, j, k \in M; g_{jk} = 0 \text{ otherwise.}$$

If non-rejected hypotheses remain, return to Step 1 and retest these hypotheses; otherwise, testing stops.

### Truncated Hommel procedure

The truncated Hommel procedure was introduced in Section 5.6.3 and can be written in the form of a step-up testing algorithm. To define this algorithm, consider a set of null hypotheses  $H_1, \dots, H_m$  with  $p_1, \dots, p_m$  denoting the corresponding raw  $p$ -values. The ordered  $p$ -values and associated ordered hypotheses are denoted by  $p_{(1)} < \dots < p_{(m)}$  and  $H_{(1)}, \dots, H_{(m)}$ , respectively. Further, let  $\gamma$  ( $0 \leq \gamma \leq 1$ ) be the truncation parameter of the truncated Hommel procedure and let  $\alpha$  be the pre-specified FWER.

The truncated Hommel procedure can be implemented using the following step-up testing algorithm which begins with  $H_{(m)}$ :

- Step  $i = 1$ . Accept  $H_{(m)}$  if  $p_{(m)} > [\gamma + (1 - \gamma)/m]\alpha$ . If  $H_{(m)}$  is accepted, proceed to Step 2. Otherwise testing stops and all null hypotheses are rejected.
- Steps  $i = 2, \dots, m - 1$ . Accept  $H_{(m-i+1)}$  and go to Step  $i + 1$  if  $p_{(m-i+j)} > [j\gamma/i + (1 - \gamma)/m]\alpha$  for all  $j = 1, \dots, i$ . Otherwise testing stops and all remaining null hypotheses  $H_{(j)}$ ,  $j = 1, \dots, m - i + 1$ , are rejected provided  $p_{(j)} \leq [\gamma/(i - 1) + (1 - \gamma)/m]\alpha$ .
- Step  $i = m$ . Accept  $H_{(1)}$  if  $p_{(j)} > [j\gamma/m + (1 - \gamma)/m]\alpha$  for all  $j = 1, \dots, m$  or  $p_{(1)} > [\gamma/(m - 1) + (1 - \gamma)/m]\alpha$ . Otherwise  $H_{(1)}$  is rejected.

The truncated Hommel procedure is separable (Dmitrienko et al., 2008) if the truncation parameter  $\gamma$  is strictly less than 1. If  $\gamma = 1$ , the truncated Hommel procedure simplifies to the regular Hommel procedure.

## Hommel-based gatekeeping procedure using the standard mixture method

Using the notation introduced in Section 5.6.2, the standard mixture method is applied to set up the Hommel-based gatekeeping procedure in the setting of Case study 6 as shown below. Let  $H_I$ ,  $I \subseteq N = \{1, \dots, 9\}$ , be an arbitrary intersection hypothesis from the closed family. The intersection  $p$ -value for  $H(I)$  is given by

$$p(I) = \min \left[ \frac{p_1(I_1^*)}{c_1(I)}, \frac{p_2(I_2^*)}{c_2(I)}, \frac{p_3(I_3^*)}{c_3(I)} \right],$$

where  $I_i^*$ ,  $i = 1, 2, 3$ , are the restricted intersection index sets with  $I_1^* = I_1$ ,  $p_i(I_i^*)$ , are  $i = 1, 2, 3$ , are the component  $p$ -values, and  $c_i(I)$ ,  $i = 1, 2, 3$ , are the predefined quantities (family weights) that defines the fraction of the overall FWER applied to each family.

The component  $p$ -values are calculated using the truncated Hommel (for Families  $F_1$  and  $F_2$  with  $\gamma_1$  and  $\gamma_2$  being the corresponding truncation parameters) or regular Hommel (for Family  $F_3$ ) procedures. Let  $p_{i(1)} < \dots < p_{i(k_i^*)}$  denote the ordered  $p$ -values within the index set  $I_i^*$ , where  $k_i^*$  denotes the number of indices in  $I_i^*$ . Further, let  $n_i$  be the number of hypotheses in Family  $F_i$ ,  $i = 1, 2, 3$ .

The component  $p$ -values are defined as follows:

$$p_i(I_i^*) = \min_{j=1, \dots, k_i^*} \frac{p_{i(j)}}{j\gamma_i/k_i^* + (1 - \gamma_i)/n_i}, \quad i = 1, 2, 3.$$

If the restricted index set  $I_i^*$  is empty, the component  $p$ -value is simply set to 1.

The family weights are defined as follows:

$$c_1(I) = 1, \quad c_2(I) = 1 - f_1(k_1, n_1 | \gamma_1), \quad c_3(I) = c_2^S(I)[1 - f_2(k_2, n_2 | \gamma_2)].$$

where  $k_i$  is the number of indices in  $I_i$  and  $f_i(k_i, n_i | \gamma_i)$  is the error fraction function of the truncated Hommel procedure. This function is given by

$$f_i(k_i, n_i | \gamma_i) = \gamma_i + \frac{(1 - \gamma_i)k_i}{n_i}$$

if  $I_i$  is non-empty ( $k_i > 0$ ) and  $f_i(k_i, n_i | \gamma_i) = 0$  otherwise. It is easy to see that the family weights are given by

$$c_2(I) = \frac{(1 - \gamma_1)(n_1 - k_1)}{n_1} \text{ if } k_1 > 0 \text{ and } c_2(I) = 1 \text{ otherwise,}$$

$$c_3(I) = c_2(I) \frac{(1 - \gamma_2)(n_2 - k_2)}{n_2} \text{ if } k_2 > 0 \text{ and } c_3(I) = c_2(I) \text{ otherwise.}$$

To illustrate the process of defining intersection  $p$ -values in the Hommel-based gatekeeping procedure, we will use the intersection hypothesis  $H_2 \cap H_4 \cap H_5 \cap H_9$ . It is easy to verify that  $I = \{2, 4, 5, 9\}$ ,  $I_1 = \{2\}$ ,  $I_2 = \{4, 5\}$ , and  $I_3 = \{9\}$ . It follows from the serial logical restrictions that the null hypothesis  $H_5$  is not testable if the null hypothesis  $H_2$  is accepted. Thus,  $H_5$  should be removed from  $I_2^*$ . When the index set  $I_3$  is considered, no changes need to be made since the null hypothesis  $H_9$  is testable if  $H_2$ ,  $H_4$ , and  $H_5$  are all accepted. This implies that  $I_1^* = \{2\}$ ,  $I_2^* = \{4\}$ ,  $I_3^* = \{9\}$ , and  $I^* = \{2, 4, 9\}$ . Thus,  $k_1 = 1$ ,  $k_2 = 2$ ,  $k_3 = 1$ , and  $k_1^* = k_2^* = k_3^* = 1$ . Since  $k_1^* = 1$  and  $n_1 = 3$ , the  $p$ -value for the component procedure used in Family  $F_1$  is

$$p_1(I_1^*) = \frac{p_2}{\gamma_1 + (1 - \gamma_1)/3} = \frac{3p_2}{1 + 2\gamma_1}.$$

Similarly,

$$p_2(I_2^*) = \frac{p_4}{\gamma_2 + (1 - \gamma_2)/3} = \frac{3p_4}{1 + 2\gamma_2}.$$

To define the component  $p$ -value in the third family, note that the truncation parameter in the regular Hommel procedure is equal to 1, which means that

$$p_3(I_3^*) = \frac{p_9}{1} = p_9.$$

The family weights for the selected intersection hypothesis are given by

$$c_1(I) = 1, \quad c_2(I) = \frac{2(1 - \gamma_1)}{3}, \quad c_3(I) = \frac{2(1 - \gamma_1)}{3} \cdot \frac{1 - \gamma_2}{3} = \frac{2(1 - \gamma_1)(1 - \gamma_2)}{9}$$

since  $k_1 = 1$ ,  $k_2 = 2$  and  $n_1 = n_2 = 3$ .

As a result, the intersection  $p$ -value for  $H_2 \cap H_4 \cap H_5 \cap H_9$  is equal to

$$p(I) = \min \left( \frac{3p_2}{1 + 2\gamma_1}, \frac{9p_4}{2(1 - \gamma_1)(1 + 2\gamma_2)}, \frac{9p_9}{2(1 - \gamma_1)(1 - \gamma_2)} \right)$$

and the intersection hypothesis is rejected if  $p(I) \leq \alpha$ .

As in Section 5.6.2, the adjusted  $p$ -values for the null hypotheses  $H_1, \dots, H_9$  are computed based on the intersection  $p$ -values, i.e.,  $\tilde{p}_i = \max p(I)$ ,  $i = 1, \dots, 9$ . The maximum is found over all index sets  $I \subseteq N$  such that  $i \in I$ .

## Hommel-based gatekeeping procedure using the modified mixture method

The modified mixture method is defined below using the Hommel-based gatekeeping procedure in Case study 6. Using the same notation as above, let  $N_i$  be the index set of the null hypotheses contained in  $F_i$ ,  $i = 1, 2, 3$ . Consider an arbitrary intersection hypothesis  $H(I)$ ,  $I \subseteq N = \{1, \dots, 9\}$ . Within the modified mixture method, the intersection  $p$ -value for  $H(I)$  is defined as

$$p^M(I) = \min \left[ \frac{p_1(I_1^*|N_1^*)}{c_1^M(I^*)}, \frac{p_2(I_2^*|N_2^*)}{c_2^M(I^*)}, \frac{p_3(I_3^*|N_3^*)}{c_3^M(I^*)} \right].$$

The superscript  $M$  is used here to denote the modified family weights, i.e.,  $c_1^M(I^*)$ ,  $c_2^M(I^*)$  and  $c_3^M(I^*)$ , and differentiate them from the family weights used in the standard method.

The restricted intersection index sets  $I_2^*$  and  $I_3^*$  used in the definition of the intersection  $p$ -value are defined as in Section 5.6.2. That is, they account for the clinically relevant logical restrictions. An important feature of the modified mixture method is that it also incorporates the logical restrictions into the definition of the family index sets. This is accomplished by defining the restricted family index sets  $N_i^*$ ,  $i = 1, 2, 3$ . When computing each component  $p$ -value, i.e.,  $p_i(I_i^*|N_i^*)$ ,  $i = 1, 2, 3$ , the corresponding component procedure is applied only to the testable null hypotheses.

To construct the restricted family index sets, note first that  $N_1^* = N_1$  since no logical restrictions are imposed in the first family. Next,  $N_i^*$  is the subset of  $N_i$  which consists only of the null hypotheses that are testable if all null hypotheses  $H_j$ ,  $j \in I_1 \cup \dots \cup I_{i-1}$ , are accepted,  $i = 2, 3$ . Let  $n_i^*$  denote the number of indices in  $N_i^*$ ,  $i = 1, 2, 3$ .

Given the restricted family index sets, the component  $p$ -values for the intersection hypothesis  $H(I)$  in the Hommel-based gatekeeping procedure are defined as follows:

$$p_i(I_i^*|N_i^*) = \min_{j=1,\dots,k_i^*} \frac{p_{i(j)}}{j\gamma_i/k_i^* + (1-\gamma_i)/n_i^*}, \quad i = 1, 2, 3,$$

where, as in the standard mixture method,  $p_{i(1)}, \dots, p_{i(k_i^*)}$  are the ordered  $p$ -values within the index set  $I_i^*$ , and  $k_i^*$  denotes the number of indices in  $I_i^*$ . An important difference between the modified and standard mixture methods is that the total number of null hypotheses in Family  $F_i$  used in the computation of the component  $p$ -value  $p_i(I_i^*|N_i^*)$  is defined as the number of testable hypotheses (i.e.,  $n_i^*$ ) rather than the actual number of hypotheses in the family (i.e.,  $n_i$ ). This approach is more relevant when the hypotheses are logically related to each other.

Another key difference is that the modified family weights are based on the restricted index set  $I^*$  rather than the original index set  $I$  and thus directly account for the logical restrictions. In particular,  $c_1^M(I^*) = 1$  and

$$c_2^M(I^*) = 1 - f_1(k_1^*, n_1^*|\gamma_1), \quad c_3^M(I^*) = c_2^M(I^*)[1 - f_2(k_2^*, n_2^*|\gamma_2)].$$

The weights used in Families  $F_2$  and  $F_3$  are given by

$$c_2^M(I^*) = \frac{(1-\gamma_1)(n_1^* - k_1^*)}{n_1^*} \text{ if } k_1^* > 0 \text{ and } c_2^M(I^*) = 1 \text{ otherwise,}$$

$$c_3^M(I^*) = c_2^M(I^*) \cdot \frac{(1-\gamma_2)(n_2^* - k_2^*)}{n_2^*} \text{ if } k_2^* > 0 \text{ and } c_3^M(I^*) = c_2^M(I^*) \text{ otherwise.}$$

To demonstrate the key features of the modified mixture method, the intersection  $p$ -value will be computed for the intersection hypothesis considered above, i.e., for  $H(I)$  with  $I = \{2, 4, 5, 9\}$ . The restricted index sets are  $I_1^* = \{2\}$ ,  $I_2^* = \{4\}$ , and  $I_3^* = \{9\}$ , which means that  $k_1 = 1$ ,  $k_2 = 2$ ,  $k_3 = 1$ , and  $k_1^* = k_2^* = k_3^* = 1$ . As pointed out above, the modified mixture method applies the logical restrictions to the family index sets. Considering  $F_1$ , we have  $N_1^* = \{1, 2, 3\}$  since no logical restrictions are imposed in this family. Further, index set  $N_2^*$  is defined by excluding the null hypotheses that become untestable if the only null hypothesis in  $I_1$  is accepted. This means that the null hypothesis  $H_5$  is removed from index set  $N_2$ . Therefore,  $N_2^* = \{4, 6\}$ . In the last family, the null hypotheses  $H_7$  and  $H_8$  depend on the null hypotheses in  $I_1 \cup I_2 = \{2, 4, 5\}$ . Thus, they need to be removed from the index set  $N_3$ , which implies that  $N_3^* = \{9\}$ . This implies that  $n_1^* = 3$ ,  $n_2^* = 2$ , and  $n_3^* = 1$ .

The intersection  $p$ -value for the selected intersection hypothesis is computed as shown below. Since the family index set in the first family is not modified,  $n_1^* = n_1 = 3$ . Therefore, the component  $p$ -value in Family  $F_1$  is found as in the standard mixture method, i.e.,

$$p_1(I_1^*|N_1^*) = \frac{p_2}{\gamma_1 + (1-\gamma_1)/3} = \frac{3p_2}{1+2\gamma_1}.$$

To obtain the component  $p$ -value in the second family, recall that only two null hypotheses are testable in this family ( $n_2^* = 2$ ), and therefore

$$p_2(I_2^*|N_2^*) = \frac{p_4}{\gamma_2 + (1-\gamma_2)/2} = \frac{2p_4}{1+\gamma_2}.$$

Given that the regular Hommel procedure is applied in the last family, it is easy to verify that the component  $p$ -value in this family is the same as in the standard mixture method, i.e.,

$$p_3(I_3^*|N_3^*) = p_9.$$

Since  $k_1^* = k_2^* = 1$ ,  $n_1^* = 3$ , and  $n_2^* = 2$ , the family weights based on the restricted index set  $I^* = \{2, 4, 9\}$  are defined as follows:

$$c_2^M(I^*) = \frac{2(1 - \gamma_1)}{3}, \quad c_3^M(I^*) = \frac{2(1 - \gamma_1)}{3} \cdot \frac{1 - \gamma_2}{2} = \frac{(1 - \gamma_1)(1 - \gamma_2)}{3}.$$

The modified intersection  $p$ -value for the selected intersection hypothesis is equal to

$$p^M(I) = \min \left( \frac{3p_2}{1 + 2\gamma_1}, \frac{3p_4}{(1 - \gamma_1)(1 + \gamma_2)}, \frac{3p_9}{(1 - \gamma_1)(1 - \gamma_2)} \right).$$

It can be shown that the modified intersection  $p$ -value for the selected intersection hypothesis is at least as significant as the intersection  $p$ -value produced by the standard method, which results in improved power for this intersection test.

As before, the adjusted  $p$ -values for the original null hypotheses  $H_1, \dots, H_9$  are given by  $\tilde{p}_i^M = \max p^M(I)$ ,  $i = 1, \dots, 9$  with the maximum computed over all index sets  $I \subseteq N$  such that  $i \in I$ .



# Chapter 6

## Interim Data Monitoring

Alex Dmitrienko (Mediana)

Yang Yuan (SAS Institute)

6.1	Introduction	251
6.2	Repeated significance tests	253
6.3	Stochastic curtailment tests	292
6.4	References	315

The chapter reviews sequential data monitoring strategies in clinical trials. It introduces a class of group sequential tests (known as repeated significance tests) that are widely used in the assessment of interim findings in late-stage trials. The first part of the chapter provides a detailed review of the process of designing group sequential trials and also discusses flexible procedures for monitoring clinical trial data. The second part of the chapter reviews popular approaches to constructing futility testing procedures in clinical trials. These approaches are based on frequentist (conditional power), mixed Bayesian-frequentist (predictive power), and fully Bayesian (predictive probability) methods.

### 6.1 Introduction

---

Sequential monitoring of safety and efficacy data has become an integral part of modern clinical trials. Although, in theory, we can consider continuous monitoring strategies, practical considerations dictate that interim monitoring be performed in a *group sequential* manner, i.e., interim looks should be taken after groups of patients have completed the trial. Within a sequential testing framework, data-driven decision rules are applied at each interim look, and the clinical trial is stopped as soon as enough information is accumulated to reach a conclusion about the properties of a new treatment, e.g., the experimental treatment is superior or inferior to the control.

In general, interim assessments of efficacy and safety data in clinical trials are motivated by the following considerations (Enas et al., 1989; Jennison and Turnbull, 1990; Ellenberg, Fleming, and DeMets, 2002):

- **Ethical requirements.** It is imperative to ensure that patients are not exposed to harmful therapies. Thus, a clinical trial must be stopped as soon as the experimental therapy is found to cause serious side effects. Interim safety evaluations are generally mandated in clinical trials with non-reversible outcomes, e.g., mortality trials.
- **Financial considerations.** In order to make the optimal use of research and development dollars, clinical trial sponsors often introduce interim analyses of efficacy endpoints in early proof-of-concept studies (as well as larger Phase II and III trials) to help predict the final outcome of the study. A decision to terminate

the study might be reached if it becomes evident that the study is unlikely to achieve its objectives at the planned end, e.g., there is very little evidence that the treatment will improve the patients' condition.

- **Administrative issues.** Clinical researchers often rely on interim data to judge the overall quality of the data and facilitate administrative or business decisions. For example, early evidence of efficacy might trigger a decision to increase manufacturing spending in order to support continuing development of the experimental treatment. The trial might still be continued to help better characterize the efficacy and safety profiles of the treatment.

The general topic of clinical trials with data-driven decision rules, known as adaptive trials, has attracted much attention across the clinical trial community over the past 15-20 years. Adaptive designs used in confirmatory Phase III clinical trials might use one or more complex decision rules aimed at increasing the total sample size in the trial or selecting the most promising treatment. The U.S. and European regulatory agencies have released several guidance documents that deal with adaptive trials:

- Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design (EMA, 2007).
- Draft Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics (FDA, 2010).
- Guidance for Industry and Food and Drug Administration Staff: Adaptive Designs for Medical Device Clinical Studies (FDA, 2016).

Dozens of research papers have been published to define statistical methods for designing, conducting, and analyzing adaptive trials and dozens of clinical trial papers have summarized the results of trials that employed adaptive designs. For an informative review of adaptive design experiences over the past 25 years, see Bauer et al. (2016).

Despite the popularity of adaptive approaches to trial design, group sequential designs with fairly simple efficacy and futility monitoring rules are still the most common types of adaptive designs. This chapter will provide a high-level summary of statistical issues arising in group sequential trials in a confirmatory setting. We will focus on the implementation of group sequential trials using SAS software as well as practical considerations related to efficacy and futility monitoring. For a detailed overview of relevant statistical theory, the reader is referred to Jennison and Turnbull (2000); Proschan, Lan, and Wittes (2006); and Wassmer and Brannath (2016).

## Overview

---

Section 6.2 provides a detailed review of group sequential designs aimed at evaluating early evidence of efficacy at one or more interim looks, including the popular O'Brien-Fleming and Pocock designs. We will show how to calculate sample size and stopping boundaries in group sequential trials based on the powerful error spending approach (Lan and DeMets, 1983). Inferences at the end of a group sequential trial will be discussed, and we will demonstrate how to calculate repeated confidence intervals for the treatment difference at each interim look as well as bias-adjusted point estimates and confidence intervals at the last analysis. PROC SEQDESIGN and PROC SEQTEST will be used to design and analyze group sequential trials.

Section 6.3 introduces an alternative approach to interim monitoring of clinical trials known as the *stochastic curtailment approach*. Stochastic curtailment methods will be applied to develop futility stopping rules for determining whether it is fruitful to continue the trial to the planned end. The section introduces three families of stochastic curtailment tests:

- Conditional power tests (frequentist approach).
- Predictive power tests (mixed Bayesian-frequentist approach).
- Predictive probability tests (Bayesian approach).

with applications to clinical trials with normally distributed and binary primary endpoints. We will show how to implement commonly used stochastic curtailment tests using PROC SEQDESIGN and PROC SEQTEST as well as SAS macros written by the authors.

Finally, it is important to remind the reader that this chapter deals with *statistical* rules for examining efficacy and futility in group sequential clinical trials. Numerous *operational* considerations that play an important role in the decision making process cannot be fully addressed within the statistical framework. It is widely recognized that a data monitoring committee (DMC) is authorized to terminate a clinical trial or modify its design for a variety of reasons, e.g., safety concerns, secondary findings, consistency of results across subsets, and convincing findings from similarly designed studies. See Ellenberg, Fleming and DeMets (2002, Chapter 8) for more details. As stated by DeMets and Lan (1984), “Just as good statistical inference is more than computing a  $p$ -value, so also is decision making in terminating a clinical trial.” For a detailed discussion of implementation and operational issues arising in clinical trials with data-driven decisions as well as more information on data monitoring committees, see Gallo, DeMets, and LaVange (2014) and Danielson et al. (2014).

The SAS code and data sets included in this chapter are available on the book’s website at <http://support.sas.com/publishing/authors/dmitrienko.html>.

## 6.2 Repeated significance tests

---

A variety of methods have been proposed in the literature for designing group sequential trials and monitoring trial data to help detect early evidence of an overwhelming treatment effect or lack of therapeutic benefit. The most popular ones are the repeated significance and boundaries approaches:

- Repeated significance testing goes back to group sequential procedures with pre-specified equally spaced looks (Pocock, 1977; O’Brien and Fleming, 1979), which were subsequently generalized to allow flexible sequential monitoring using the error spending approach proposed by Lan and DeMets (1983).
- The boundaries approach (triangular and related tests) represents an extension of sequential procedures for continuous data monitoring (e.g., Wald’s sequential probability ratio test), which were suitably modified to be used in a group sequential manner.

Repeated significance testing is currently the most widely used approach and will be the focus of this section. For a comprehensive review of the boundaries approach, see Whitehead (1997) and other publications by John Whitehead and his colleagues (for example, Whitehead and Stratton (1983) and Whitehead (1999)).

### 6.2.1 Group sequential trial designs

To illustrate group sequential designs based on repeated significance tests, consider a clinical trial with a balanced design that compares the efficacy profile of an experimental treatment to a control using a normally distributed primary endpoint. As always, the case of normally distributed outcome variables is generally applicable to the analysis of test statistics computed from non-normal data. Group sequential designs for binary or time-to-event endpoints are conceptually similar to those discussed below<sup>1</sup>.

Assume that the following group sequential design with  $m$  interim analyses will be adopted in the trial. Let  $N$  denote the maximum sample size per trial arm. Further,  $X_{i1}, \dots, X_{iN}$  will denote the independent, identically distributed measurements in the  $i$ th group,  $i = 1, 2$ . The treatment effect  $\delta$  is equal to the difference between the treatment means. For simplicity, suppose that  $2n$  patients complete the trial between successive looks ( $n$  patients per arm). The null hypothesis of no treatment effect will be tested at the  $k$ th interim analysis using the test statistic

$$Z_k = \sqrt{\frac{kn}{2s^2}} \left( \frac{1}{kn} \sum_{j=1}^{kn} X_{1j} - \frac{1}{kn} \sum_{j=1}^{kn} X_{2j} \right),$$

where  $s$  denotes the pooled sample standard deviation.

The objective of the trial is to test the null hypothesis of no treatment effect

$$H_0 : \quad \delta = \delta_0 = 0$$

against a one-sided alternative

$$H_1 : \quad \delta = \delta_1 > 0.$$

Here  $\delta_1$  denotes a clinically meaningful treatment difference.

We will now consider the problem of setting up group sequential designs for evaluating early evidence of superior efficacy or futility. This is typically accomplished using one-sided tests or stopping boundaries. It is worth noting that group sequential tests introduced below are easily generalized to a two-sided setting. This is accomplished by adding symmetric stopping boundaries. For example, a group sequential design with a two-sided alternative relies on two symmetric boundaries. Similarly, simultaneous efficacy/futility testing with a two-sided alternative hypothesis requires two sets of two symmetric boundaries.

Also, the main focus of Section 6.2 is standard group sequential designs aimed at evaluating early evidence of efficacy. Clinical trials with futility stopping rules will be discussed in Section 6.3 in the context of alternative data monitoring methods (stochastic curtailment tests).

### Group sequential designs for detecting superior efficacy

The null hypothesis is tested by sequentially comparing the test statistics  $Z_1, \dots, Z_m$  to adjusted critical values  $u_1(\alpha), \dots, u_m(\alpha)$  forming an upper stopping boundary.

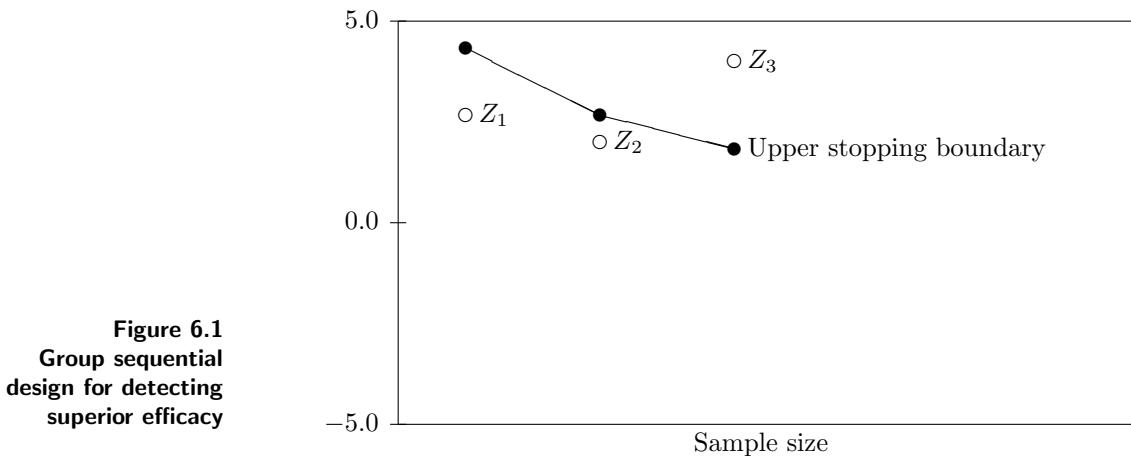
---

<sup>1</sup>As shown by Jennison and Turnbull (2000, Chapter 13), the group sequential designs and sequential monitoring strategies discussed later in this section are easily extended to the case of time-to-event variables. An interesting feature of group sequential designs (in trials with time-to-event endpoints that differentiates them from normally distributed and binary endpoints) is a repeated use of data collected from the same patient. For example, in survival trials, the test statistic at each interim look is computed from all survival times (either observed or censored); thus, any patient who survives to the end of the study contributes to every interim test statistic.

The adjusted critical values  $u_1(\alpha), \dots, u_m(\alpha)$  are selected to preserve the overall Type I error rate  $\alpha$ . The following decision rule is used at the  $k$ th interim look:

1. Stop the trial and conclude that the experimental treatment is superior to placebo if  $Z_k > u_k(\alpha)$ ;
2. Continue the trial if  $Z_k \leq u_k(\alpha)$ .

In other words, a clinical trial designed to study the efficacy of a novel treatment is discontinued when the test statistic crosses the upper stopping boundary. For example, Figure 6.1 depicts the results observed in a hypothetical clinical trial, which is stopped at the third interim look due to overwhelming evidence of therapeutic benefit.



### Group sequential designs for detecting futility

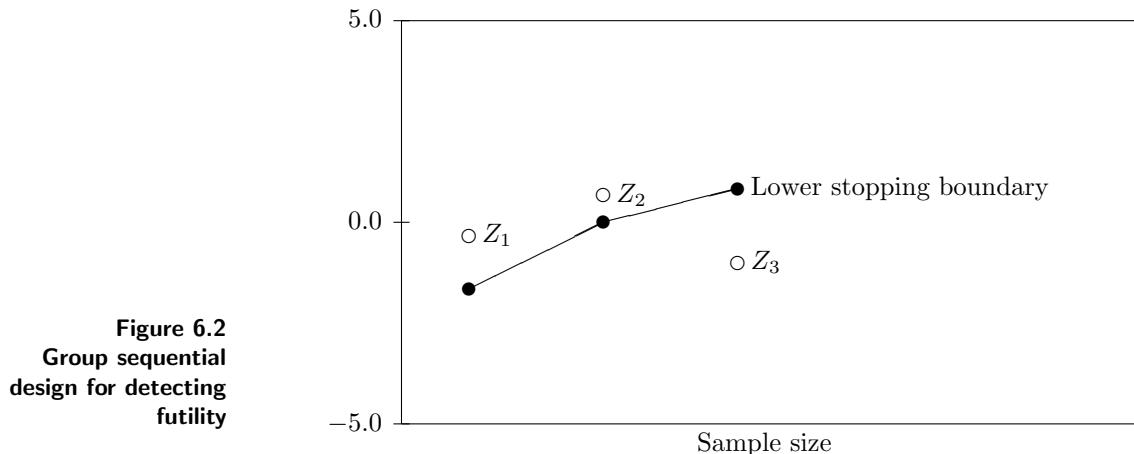
Even though it is rarely done in practice, the general group sequential testing approach can be applied to define tests for detecting futility. In this case, futility testing will rely on comparing the statistics  $Z_1, \dots, Z_m$  to critical values  $l_1(\beta), \dots, l_m(\beta)$  forming a lower stopping boundary. (The critical values are chosen to maintain the nominal Type II error rate.) The experimental treatment is declared futile at the  $k$ th interim analysis if the accumulated evidence suggests that the alternative hypothesis is false, i.e.,

$$Z_k < l_k(\beta),$$

and continued otherwise. To illustrate this decision rule, Figure 6.2 presents a scenario in which a clinical trial is terminated at the third interim look due to futility.

### Group sequential designs for simultaneous efficacy and futility testing

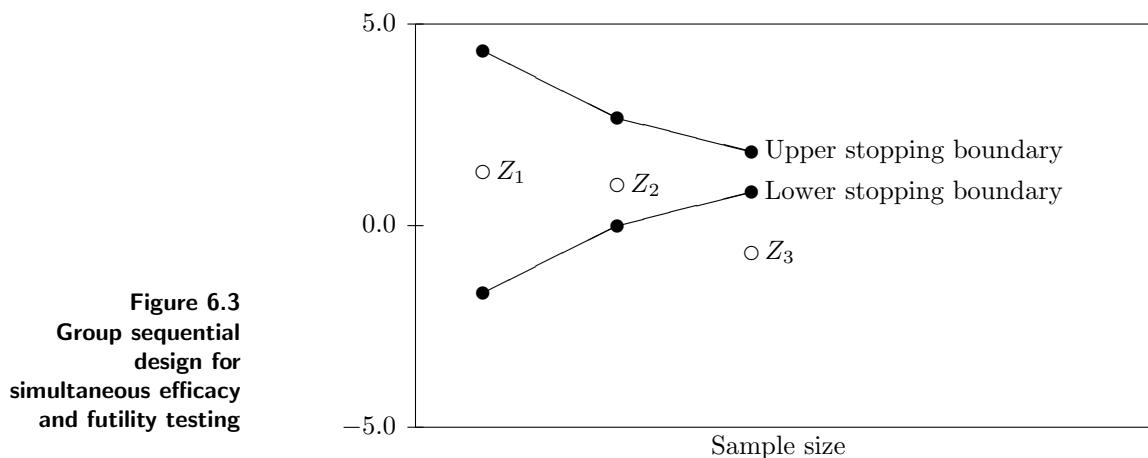
In addition, the group sequential testing approach can be used to define trial designs for simultaneously assessing the efficacy and futility profiles of experimental treatments. A trial with this design will be stopped when the test statistic crosses



either the lower or upper boundary. In this case, the decision rule at the  $k$ th interim analysis looks like this:

1. Stop the trial for efficacy if  $Z_k > u_k(\alpha, \beta)$ ;
2. Continue the trial if  $l_k(\alpha, \beta) \leq Z_k \leq u_k(\alpha, \beta)$ ;
3. Stop the trial for futility if  $Z_k < l_k(\alpha, \beta)$ .

Figure 6.3 provides an example of a clinical trial terminated at the third interim look because the test statistic crossed the lower stopping boundary.



It is important to note that the stopping boundaries in trial designs for simultaneous efficacy and futility testing must be selected to protect both the Type I and II error probabilities  $\alpha$  and  $\beta$ , which is rather uncommon in clinical trials. As explained in Section 6.3, futility stopping rules are typically applied in a *non-binding* manner in the sense that the trial's sponsor reserves the right to override a decision to terminate the trial due to apparent lack of efficacy. To maintain  $\alpha$  control with non-binding futility stopping rules, the upper stopping boundary is computed first to protect the Type I error rate, and an appropriate futility stopping rule is constructed separately from this efficacy stopping boundary. The resulting set of decision rules no longer controls the Type II error rate.

## 6.2.2 Case studies

The following two case studies based on clinical trials with normally distributed or binary endpoints will be used to illustrate the design and analysis of group sequential trials that support early stopping due to superior efficacy.

**EXAMPLE: Case study 1 (Trial in patients with clinical depression)**

A two-arm clinical trial was conducted to test a single dose of a novel treatment for treating disorders of the central nervous system. The trial's primary objective was to demonstrate that the chosen dose of the treatment was superior to placebo in the acute treatment of subjects who meet criteria for major depression. The efficacy of the experimental treatment was evaluated using the mean change from baseline to the end of the 8-week study period in the total score of the 17-item Hamilton Depression Rating Scale (HAMD17). A lower HAMD17 score indicated a beneficial effect.

To determine the size of the trial, the sponsor wanted to have adequate power to detect a treatment difference in the mean reduction in the HAMD17 total score when this difference is 3 with a standard deviation of 8. With a traditional design, a sample size of 150 patients per trial arm provides a 90% power to detect a statistically significant difference at a two-sided 0.05 level.

The clinical trial employed two interim analyses and a final analysis. The analyses were planned to take place after approximately 50 and 75 percent of the patients completed the study. The patient enrollment was to be terminated in the presence of overwhelming efficacy at either analysis. The data collected in the trial are summarized in Table 6.1. Note that the mean change from baseline in the HAMD17 score (negative quantity) is presented as the mean improvement (positive quantity).

**TABLE 6.1 Summary of HAMD17 data collected in the depression trial**

Analysis	Treatment arm		Placebo arm	
	n	Mean improvement in HAMD17 total score (SD)	n	Mean improvement in HAMD17 total score (SD)
Interim analysis 1	78	8.3 (6.2)	78	5.9 (6.5)
Interim analysis 2	122	8.0 (6.3)	120	6.3 (5.9)
Final analysis	150	8.2 (5.9)	152	6.1 (5.8)

**EXAMPLE: Case study 2 (Trial in patients with severe sepsis)**

A placebo-controlled Phase III clinical trial was designed to study the effect of an experimental treatment on 28-day all-cause mortality in patients with severe sepsis. The objective of the trial was to assess whether the treatment had superior efficacy compared to the placebo.

It was assumed that the 28-day placebo mortality was 30% and administration of the treatment reduced the mortality rate to 24%. Using a traditional design with a fixed sample size and assuming a two-sided significance level of 0.05, the total of 1718 patients would need to be enrolled in the trial to achieve 80% power of detecting an absolute mortality reduction of 6%.

The mortality data were examined at two interim analyses in this Phase III trial. These analyses were scheduled to occur after approximately 20 and 66 percent of the patients completed the trial. The first interim analysis was introduced mainly to assess the futility of the experimental treatment. The trial was to be terminated early for efficacy only in the presence of an extremely beneficial treatment effect. The objective of the second interim analysis was to examine both efficacy and futility.

The 28-day survival data collected at the two interim analyses are presented in Table 6.2.

**TABLE 6.2 Summary of 28-day survival data collected in the sepsis trial**

Analysis	Treatment arm			Placebo arm		
	n	Number of survivors	Survival rate	n	Number of survivors	Survival rate
Interim analysis 1	220	164	74.5%	216	152	70.4%
Interim analysis 2	715	538	75.2%	715	526	73.6%

### 6.2.3 Design and monitoring stages

When considering available group sequential designs, it is helpful to draw a line between the design and analysis (or monitoring) stages. The two components of a group sequential trial are supported in SAS using PROC SEQDESIGN and PROC SEQTEST, respectively. The following is a brief description of steps we need to go through to design a group sequential trial and perform interim monitoring.

#### Design stage (PROC SEQDESIGN)

- Specify the stopping boundaries in a group sequential trial based on a fixed or flexible data monitoring approach, e.g., using a pre-defined set of decision points or an error spending function. The choice of stopping boundaries is driven by a variety of factors that will be discussed in Section 6.2.4.
- Design a group sequential trial and compute its key operating characteristics such as the maximum number of patients, average sample size and power as a function of the treatment difference. See Sections 6.2.5, 6.2.6 and 6.2.7.

#### Monitoring stage (PROC SEQTEST)

- Implement a flexible data monitoring scheme based on the predefined error spending function. Using the error spending function, compute adjusted critical values and *p*-value cutoff points at each interim analysis, see Section 6.2.8. The advantage of the error spending approach is that it provides the trial's sponsor or independent data monitoring committee with much flexibility and enables them to deviate from the prespecified sequential plan in terms of the number or timing of interim looks.
- Compute repeated confidence limits for the true treatment difference at each interim look (Section 6.2.9) and a bias-adjusted estimate of the treatment difference and confidence limits at the last analysis (Section 6.2.10).

### 6.2.4 Fixed and flexible data monitoring strategies

There are two main approaches to designing and conducting clinical trials with data monitoring schemes. The first one relies on a set of prespecified time points at which the data are to be reviewed. The Pocock and O'Brien-Fleming stopping boundaries (Pocock, 1977; O'Brien and Fleming, 1979) were originally proposed for this type of inspection scheme. An alternative approach was introduced by Lan and DeMets (1983). It is based on a flexible *error spending* strategy and enables the trial's sponsor or data monitoring committee to change the timing and frequency of

interim looks. Within the error spending framework, interim monitoring follows the philosophical rationale of the design thinking while allowing considerable flexibility. These fixed and flexible data monitoring strategies are described in this section.

### Fixed strategies

Group sequential design for detecting superior efficacy is often set up using stopping boundaries proposed by Pocock (1977) and O'Brien and Fleming (1979). These authors were among the first to develop statistical methods for designing group sequential clinical trials. Pocock (1977) described a set of stopping boundaries with the same critical value at each interim look. O'Brien and Fleming (1979) proposed a sequential plan under which earlier analyses are performed in a conservative manner, and the later tests are carried out at significance levels close to the nominal level.

The Pocock and O'Brien-Fleming stopping boundaries for efficacy monitoring are defined in Table 6.3. The adjusted critical values (boundary values) are derived using a straightforward calculation under the assumption that the interim analyses are equally spaced. Specifically, the constants  $c_P(\alpha, m)$  and  $c_{OF}(\alpha, m)$  shown in the table are chosen to protect the overall Type I error rate, and  $m$  denotes the total number of decision points (interim and final analyses) in the trial.

**TABLE 6.3** Popular approaches to constructing stopping boundaries for efficacy testing in clinical trials with equally spaced interim looks.

Approach	Adjusted critical values
Pocock	$u_k(\alpha) = c_P(\alpha, m), k = 1, \dots, m$
O'Brien-Fleming	$u_k(\alpha) = c_{OF}(\alpha, m)k^{-1/2}, k = 1, \dots, m$

As a quick illustration, consider a clinical trial with a single interim analysis ( $m = 2$ ), and assume that the one-sided Type I error rate is  $\alpha = 0.025$ . Suppose that the information fraction at this interim look is  $t_1 = 0.5$ . In this case,  $c_P(0.025, 2) = 2.18$ , which means that the constant Pocock-adjusted critical value used at both decision points is equal to

$$u_1(0.025) = u_2(0.025) = c_P(0.025, 2) = 2.18.$$

Considering the O'Brien-Fleming approach to defining stopping boundaries,  $c_{OF}(0.025, 2) = 2.80$ . Thus, the corresponding adjusted critical values at the interim analysis and final analysis are given by

$$u_1(0.025) = c_{OF}(0.025, 2) = 2.80, \quad u_2(0.025) = c_{OF}(0.025, 2)/\sqrt{2} = 1.98.$$

### Flexible strategies

The error spending approach serves as a flexible alternative to group sequential designs with a fixed data monitoring schedule. Why does the sponsor need flexibility with respect to timing and frequency of analyses? It is sometimes convenient to tie interim looks to *calendar time* rather than *information time* related to the sample size. Non-pharmaceutical studies often employ interim analyses performed at regular intervals, e.g., every 3 or 6 months. (See, for example, Van Den Berghe et al., 2001.) Flexible strategies are also preferable in futility monitoring. From a logistical perspective, it is more convenient to perform futility analyses on a monthly or quarterly basis rather than after a prespecified number of patients have been enrolled into the trial. In this case, the number of patients changes unpredictably between looks, and we need to find a way to deal with random increments of information in the data analysis.

This section provides a high-level review of the error spending methodology. For a further discussion of the methodology, see Jennison and Turnbull (2000, Chapter 7) or Proschan, Lan and Wittes (2006, Chapter 5).

To introduce the error spending approach, consider a two-arm clinical trial with a balanced design ( $n$  patients per trial arm). The trial's sponsor is interested in implementing a group sequential design to facilitate the detection of early signs of therapeutic benefit. A Type I error spending function  $\alpha(t)$  is a non-decreasing function of the fraction of the total sample size  $t$  ( $0 \leq t \leq 1$ ), known as the *information fraction*, and

$$\alpha(0) = 0 \text{ and } \alpha(1) = \alpha,$$

where  $\alpha$  is the prespecified Type I error rate, e.g., a one-sided  $\alpha = 0.025$ . Suppose that analyses are performed after  $n_1, \dots, n_m$  patients have been accrued in each arm ( $n_m = n$  is the total number of patients per arm). It is important to emphasize that interim looks can occur at arbitrary time points. Thus,  $n_1, \dots, n_m$  are neither prespecified nor equally spaced. Let  $t_k = n_k/n$  and  $Z_k$  denote the information fraction and test statistic at the  $k$ th look, respectively. The joint distribution of the test statistics is assumed to be multivariate normal. Finally, denote the true treatment difference by  $\delta$ .

Put simply, the selected error spending function determines the rate at which the overall Type I error probability is spent during the trial. To see how it works, suppose that the first interim look is taken when the sample size in each treatment group is equal to  $n_1$  patients. An upper one-sided critical value, denoted by  $u_1$ , is determined in such a way that the amount of Type I error spent equals  $\alpha(n_1/n)$ , which is equal to  $\alpha(t_1)$ . In other words, choose  $u_1$  to satisfy the following criterion

$$P\{Z_1 > u_1 \text{ if } \delta = 0\} = \alpha(t_1).$$

The trial is stopped at the first interim analysis if  $Z_1$  is greater than  $u_1$ , and a decision to continue the trial is made otherwise.

Since we have already spent a certain fraction of the overall Type I error at the first analysis, the amount we have left for the second analysis is

$$\alpha(n_2/n) - \alpha(n_1/n) = \alpha(t_2) - \alpha(t_1).$$

Therefore, at the time of the second interim look, the critical value  $u_2$  is obtained by solving the following equation

$$P\{Z_1 \leq u_1, Z_2 > u_2 \text{ if } \delta = 0\} = \alpha(t_2) - \alpha(t_1).$$

Again, compare  $Z_2$  to  $u_2$  and proceed to the next analysis if  $Z_2$  does not exceed  $u_2$ .

Likewise, the critical value  $u_3$  used at the third interim analysis is defined in such a way that

$$P\{Z_1 \leq u_1, Z_2 \leq u_2, Z_3 > u_3 \text{ if } \delta = 0\} = \alpha(t_3) - \alpha(t_2)$$

and a similar argument is applied in order to compute how much Type I error can be spent at each of the subsequent analyses and determine the corresponding critical values  $u_4, \dots, u_m$ . It is easy to verify that the overall Type I error associated with the constructed group sequential test is equal to

$$\alpha(t_1) + [\alpha(t_2) - \alpha(t_1)] + [\alpha(t_3) - \alpha(t_2)] + \dots + [\alpha(t_m) - \alpha(t_{m-1})] = \alpha(t_m)$$

and, by the definition of an  $\alpha$ -spending function,

$$\alpha(t_m) = \alpha(1) = \alpha.$$

The described sequential monitoring strategy preserves the overall Type I error rate regardless of the timing and frequency of interim looks.

To reiterate, the Lan-DeMets error spending approach allows for flexible interim monitoring without sacrificing the overall Type I error probability. However, the power of a group sequential trial might depend on the chosen monitoring strategy. As demonstrated by Jennison and Turnbull (2000, Section 7.2), the power is generally lower than the target value when the looks are more frequent than anticipated and greater than the target value otherwise. In the extreme cases examined by Jennison and Turnbull, the attained power differed from its nominal value by about 15%.

We have focused on error spending functions for the Type I error ( $\alpha$ -spending functions). As shown by Pampallona and Tsiatis (1994) and Pampallona, Tsiatis, and Kim (2001)<sup>2</sup>, the outlined approach is easily extended to group sequential trials for futility monitoring, in which case a Type II error spending function is introduced, or simultaneous efficacy and futility monitoring, which requires a specification of both Type I and Type II error spending functions.

### Error spending functions

Multiple families of  $\alpha$ -spending functions have been introduced in the literature. See, for example, Lan and DeMets (1983), Jennison and Turnbull (1990, 2000) and Hwang, Shih, and DeCani (1990). These families are supported by both PROC SEQDESIGN and PROC SEQTEST. The Lan-DeMets family of error spending functions is defined below.

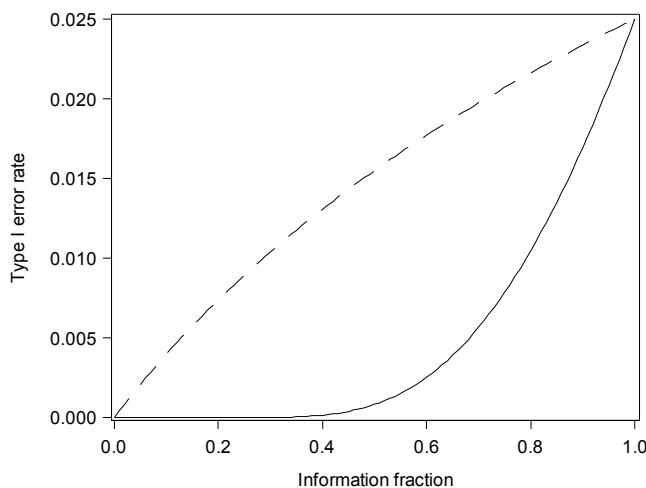
Lan and DeMets (1983) introduced two  $\alpha$ -spending functions that approximate the stopping boundaries associated with the O'Brien-Fleming and Pocock group sequential designs with equally spaced interim looks. For any prespecified overall Type I error probability  $\alpha$ , e.g., a one-sided  $\alpha = 0.025$ , the functions are given by

$$\begin{aligned}\alpha(t) &= 2 - 2\Phi(z_{1-\alpha/2}/\sqrt{t}) \text{ (O'Brien-Fleming approach),} \\ \alpha(t) &= \alpha \ln(1 + (e - 1)t) \text{ (Pocock approach),}\end{aligned}$$

where  $t$  is the information fraction and  $\Phi(x)$  is the cumulative probability function of the standard normal distribution.

The two functions are depicted in Figure 6.4. It is evident from this figure that, with the  $\alpha$ -spending function based on the O'Brien-Fleming approach, the Type

**Figure 6.4**  
Hypothetical trial example



$\alpha$ -spending functions based on the O'Brien-Fleming approach (solid curve) and Pocock approach (dashed curve).

<sup>2</sup>Although the paper by Pampallona, Tsiatis, and Kim was published in 2001, it was actually written prior to 1994.

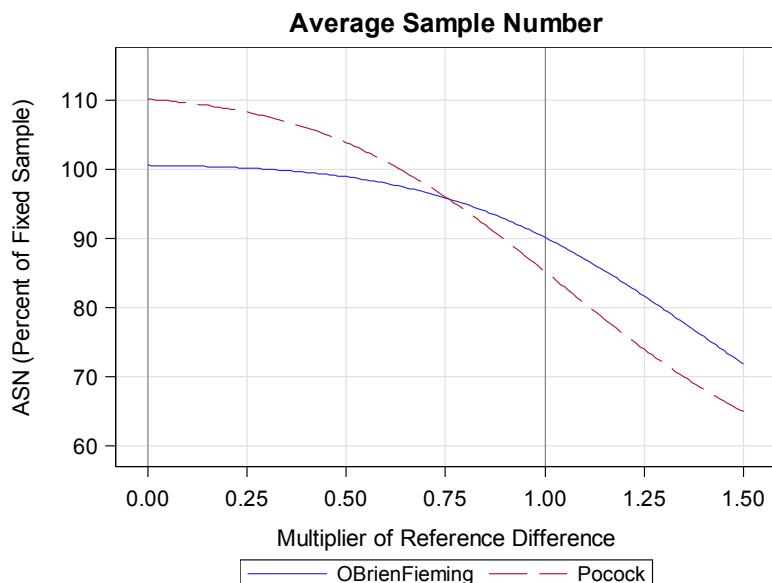
I error rate is spent very slowly early in a trial. This will lead to adjusted critical values at early interim looks that are much greater than the critical value used in traditional designs with a fixed sample size. By contrast, if the Pocock approach is considered, the Type I error rate will be spent at a linear rate, which will result in less extreme adjusted critical values at early interim analyses.

### Comparison of group sequential designs

A detailed comparison of the Pocock and O'Brien-Fleming stopping boundaries and associated  $\alpha$ -spending functions in group sequential trials will be provided in Sections 6.2.5 and 6.2.6. It will be shown that, as stated above, the chances of stopping the trial well before its planned termination are quite high if the Pocock approach is applied. By contrast, the O'Brien-Fleming design “spends” very small amounts of  $\alpha$  at early interim looks, and, as a consequence, the probability of early stopping tends to be low.

To provide a quick comparison of the Pocock and O'Brien-Fleming approaches to designing group sequential trials, consider the following hypothetical example. A two-arm clinical trial with a continuous primary endpoint has been designed to compare an experimental treatment to placebo. The trial has been sized to achieve 80% power at a two-sided 0.05 significance level under the assumption that the mean treatment difference is equal to 0.2 and the standard deviation is 1. Figures 6.5 and 6.6 show how much savings in terms of the total number of patients that we should expect with trial designs based on the two stopping boundaries. The figures plot the average number of patients to be enrolled in trials that employ the Pocock and O'Brien-Fleming group sequential designs with one and four interim looks relative to the sample size used in a traditional design with a predefined sample size. The mean treatment difference is plotted on the horizontal axis relative to the

**Figure 6.5**  
Hypothetical trial example



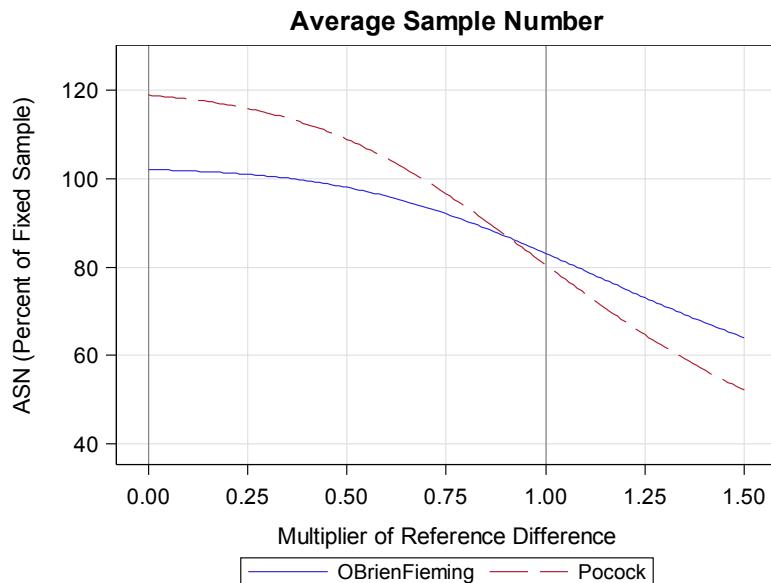
Average sample size in the O'Brien-Fleming (solid curve) and Pocock (dashed curve) group sequential designs with a single interim analysis compared to a traditional design.

reference difference of 0.2.

Figure 6.5 demonstrates that, with only one interim analysis that occurs half-way into the trial, the average number of patients required by the O'Brien-Fleming design is 10% smaller than that of a traditional design when the mean treatment difference is truly equal to 0.2 (which corresponds to the value of 1 on the horizontal axis). Assuming the same mean difference, an application of the Pocock design results in a 15% savings. The average sample size reduction increases with the true mean difference. For example, it is equal to 28% (O'Brien-Fleming design) and 35% (Pocock design) when the true mean treatment difference is 0.3. On the other hand, sequential testing can potentially increase the average sample size. For example, we can see from Figure 6.5 that the average number of patients required by the Pocock design is greater than that of the traditional design when the treatment difference is close to 0.

We should expect a more substantial reduction in the average sample size with more frequent interim looks. Figure 6.6 shows that the O'Brien-Fleming and Pocock designs with four equally spaced interim analyses offer an 18% and 20% reduction in the number of patients, respectively, compared to a traditional design when the true mean difference is equal to 0.2.

**Figure 6.6**  
Hypothetical trial example



Average sample size in the O'Brien-Fleming (solid curve) and Pocock (dashed curve) group sequential designs with four equally spaced interim analyses compared to a traditional design.

### 6.2.5 Design stage: O'Brien-Fleming approach in Case study 1

To illustrate the process of setting up a group sequential plan for detecting early evidence of superior efficacy, consider the clinical trial in patients with clinical depression introduced in Section 6.2.2. There are three decision points in this trial:

- First interim analysis after 50% of the patients complete the trial.
- Second interim analysis after 75% of the patients complete the trial.
- Final analysis.

The information fractions at these decision points are given by  $t_1 = 0.5$ ,  $t_2 = 0.75$  and  $t_3 = 1$ . Both interim analyses focus on testing superior efficacy of the experimental treatment compared to placebo.

Programs 6.1 and 6.2 demonstrate how PROC SEQDESIGN can be used to set up a group sequential design for this setting using the O'Brien-Fleming approach. First, the O'Brien-Fleming stopping boundaries will be computed in Program 6.1 using a fixed data monitoring strategy that does not rely on  $\alpha$ -spending functions. An alternative approach with an O'Brien-Fleming-type  $\alpha$ -spending function introduced in Lan and DeMets (1983) will be used in Program 6.2. Fixed and flexible data monitoring strategies are described in Section 6.2.4.

Beginning with the fixed data monitoring strategy, Program 6.1 sets up a group sequential design with an O'Brien-Fleming stopping boundary using `method=obf`. The total number of decision points in the trial (two interim looks and final analysis) is specified using `nstages=3`. The information fractions at the three decision points are given by `info=cum(0.5 0.75 1)`. The null hypothesis of no treatment effect will be rejected in the depression trial if the test statistic crosses the upper boundary and thus `alt=upper`. The Type I and Type II error rate in the trial are equal to `alpha=0.025` and `beta=0.1`. Using this information, PROC SEQDESIGN can compute the adjusted critical values at the two interim looks as well as the final analysis.

In addition, PROC SEQDESIGN can compute key characteristics of the resulting group sequential trial. This includes, for example, the probabilities of stopping at each interim analysis to claim a statistically significant treatment effect. These probabilities are typically computed under the null hypothesis of no effect and the alternative hypothesis of a beneficial treatment effect, which is requested by `stopprob(cref=0 1)`. The following option indicates that the total sample size will be shown on the horizontal axis in the stopping boundary plot and also request plots of the power function and average sample size in the trial:

```
plots=(boundary(hscale=samplesize) power asn)
```

Further, PROC SEQDESIGN can perform power calculations and compute the average number of patients in the resulting group sequential trial. To request sample size calculations, the treatment effect under the alternative hypothesis needs to be specified, i.e.,

```
samplesize model=twosamplemeans (meandiff=3 stddev=8);
```

The primary analysis in this trial relies on a comparison of two means, and, as stated in Section 6.2.2, the assumed mean treatment difference and common standard deviation are 3 and 8, respectively.

Finally, key data sets, e.g., a data set with the upper stopping boundary, can be saved using the ODS statements shown at the end of Program 6.1. This particular data set (`obf_boundary`) will be used in Section 6.2.8 to set up efficacy monitoring based on the O'Brien-Fleming group sequential design.

## **PROGRAM 6.1    Group sequential design with a fixed O'Brien-Fleming stopping boundary in Case study 1**

```
proc seqdesign stopprob(cref=0 1)
            plots=(boundary(hscale=samplesize) power asn);
OBrienFleming: design method=obf
                nstages=3
                info=cum(0.5 0.75 1)
                alt=upper
```

```

stop=reject
alpha=0.025
beta=0.1;
samplesize model=twosamplemeans (meandiff=3 stddev=8);
ods output boundary=obf_boundary;
run;

```

Program 6.1's output is organized in multiple tables. Key characteristics of the O'Brien-Fleming sequential plan in Case study 1 are presented below in Tables 6.4, 6.5 and 6.6. First of all, Table 6.4 defines the stopping boundary that will be used in the group sequential trial. As shown in the table, the adjusted critical values (boundary values) at these decision points based on the O'Brien-Fleming approach are given by

$$u_1(0.025) = 2.863, \ u_2(0.025) = 2.337, \ u_3(0.025) = 2.024.$$

As expected with the O'Brien-Fleming approach, the trial will be terminated at either interim look only if the treatment effect is highly statistically significant. For example, the test statistic needs to exceed 2.863 at the first interim analysis, and this adjusted critical value is substantially higher than the critical value used in traditional designs (1.959).

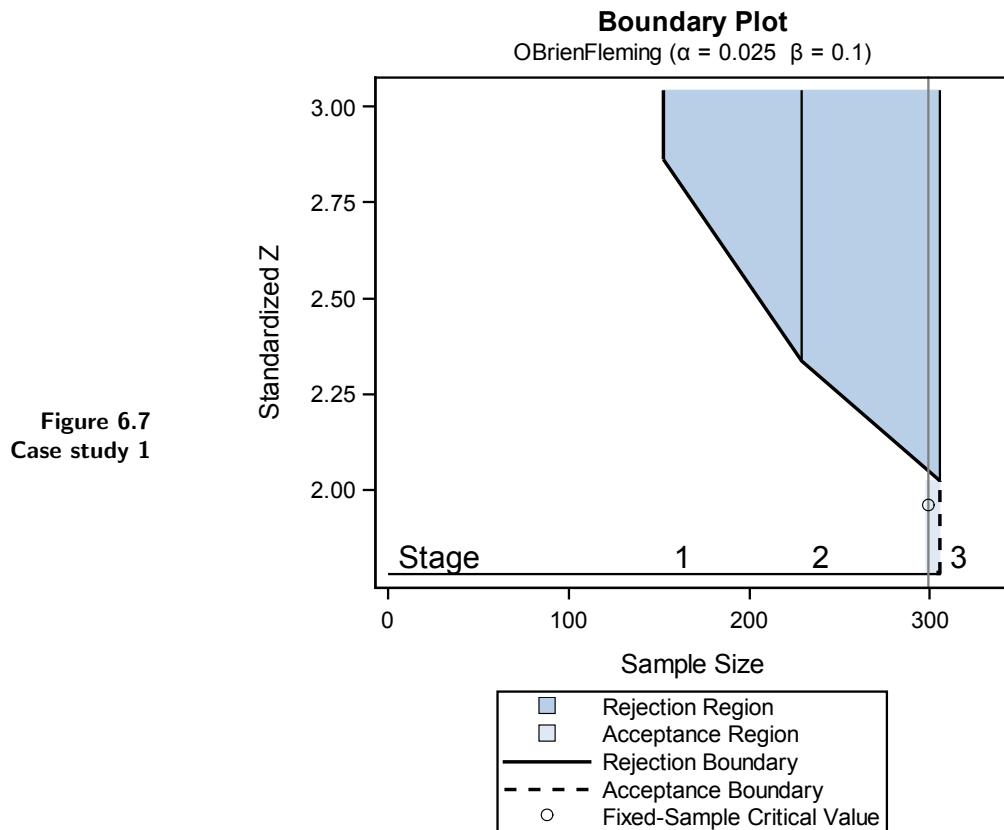
**TABLE 6.4 Output from Program 6.1: Boundary information**

Stage	Boundary Information (Standardized Z Scale)				
	Information Level			Alternative Reference	Boundary Values
	Proportion	Actual	N	Upper	Upper Alpha
1	0.50	0.597	152.7	2.317	2.863
2	0.75	0.895	229.1	2.838	2.337
3	1.00	1.193	305.5	3.277	2.024

To better understand the principles underlying the resulting group sequential design, it is helpful to plot the stopping boundary defined in Table 6.4 against the sample size per trial arm. Figure 6.7 displays the O'Brien-Fleming boundary on a test statistic (Standardized Z) scale. Although the stopping boundary appears to be continuous, it is important to remember that this is not really the case. The actual boundary is discrete. (It is represented by the lower ends of the vertical lines at the three decision points.) And the lines connecting the lower ends are introduced to help visualize the effect of repeated significant testing on the adjusted critical values. Also, to emphasize that only one analysis is performed under the traditional design with a fixed sample size of 300 patients, the traditional setting is represented by a single circle. As shown in Figure 6.7, under the O'Brien-Fleming plan, early analyses are performed in a very conservative fashion. As a result, the final analysis is conducted at a significance level close to the level used in a traditional setting. Indeed, the adjusted critical values at the first two looks are well above the open circle corresponding to the unadjusted critical value of 1.959. However, the difference between the adjusted and unadjusted critical values at the final analysis is rather small.

Using the information presented in Table 6.4, it is easy to express the O'Brien-Fleming decision rules in terms of observed mean treatment differences at each of the two interim analyses and the final analysis. The latter can be included in the data monitoring committee interim analysis guidelines (also known as the *data monitoring committee charter*). The following are examples of such guidelines:

- The first interim analysis will be conducted after approximately 77 patients have completed the study in each trial arm. The trial will be stopped for efficacy if the



*Stopping boundary of the O'Brien-Fleming design and critical value of the traditional design on a test statistic scale.*

treatment effect statistic is greater than 2.863 (i.e., the one-sided  $p$ -value is less than 0.0021). Assuming that the standard deviation of changes in the HAMD17 total score is 8, the trial will be terminated at the first interim analysis provided that the mean treatment difference exceeds 3.69 points<sup>3</sup>.

- The second interim analysis will occur after approximately 115 patients have completed the study in each arm. The trial will be stopped for efficacy if the treatment effect statistic is greater than 2.337 or, equivalently, the one-sided  $p$ -value is less than 0.0097. The trial will be terminated at the second interim analysis provided the mean treatment difference is greater than 2.47 points (under the assumption that the standard deviation of HAMD17 changes equals 8).
- The experimental treatment will be declared superior to placebo at the final analysis if the treatment effect statistic is greater than 2.024 or, equivalently, if the one-sided  $p$ -value is significant at the 0.0215 level. A significant  $p$ -value will be observed if the mean treatment difference exceeds 1.85 (assuming that the standard deviation of HAMD17 changes is 8).

This summary reminds us again that, with the O'Brien-Fleming sequential plan, the adjusted critical value at the final analysis is only slightly greater than the critical value used in a traditional trial with a fixed sample size, i.e., 1.959.

<sup>3</sup>This treatment difference is equal to  $u_1(0.025)\sigma\sqrt{2/n}$ , where  $\sigma$  is the common standard deviation and  $n$  is the sample size per arm. In this case,  $u_1(0.025) = 2.863$ ,  $\sigma = 8$  and  $n = 77$ . Thus, the detectable treatment difference is 3.69.

Table 6.5 provides general information on the group sequential design and sample size calculations. Recall from Section 6.2.2 that, with a traditional design with a single decision point at the final analysis, this trial is powered at 90% if the total sample size is set to 300 patients. It is important to study the number of patients that are expected to be enrolled in a group sequential trial in order to reach a decision point and compare it to the sample size in the corresponding traditional design. It follows from Table 6.5 that, with the group sequential design based on the O'Brien-Fleming approach, the maximum total sample size is 305.5. Although this maximum sample size is slightly greater than the sample size in the traditional setting with a fixed sample size, only 230 patients are anticipated to be enrolled in the group sequential trial under the alternative hypothesis (Alt Ref). We should on average see a 23% reduction in the sample size if the mean treatment difference is truly equal to 3. Further, if the treatment does not provide a beneficial effect (Null Ref), the expected total sample size in the group sequential trial is 304.5 patients, which is again quite close to the total sample size of 300 patients in the traditional design.

**TABLE 6.5 Output from Program 6.1: Design information and Sample size summary**

Design information	
Statistic Distribution	Normal
Boundary Scale	Standardized Z
Alternative Hypothesis	Upper
Early Stop	Reject Null
Method	O'Brien-Fleming
Boundary Key	Both
Alternative Reference	3
Number of Stages	3
Alpha	0.025
Beta	0.1
Power	0.9
Max Information (Percent of Fixed Sample)	102.2
Max Information	1.193
Null Ref ASN (Percent of Fixed Sample)	101.9
Alt Ref ASN (Percent of Fixed Sample)	76.9
Sample size summary	
Two-Sample Means	
Test	
Mean Difference	3
Standard Deviation	8
Max Sample Size	305.5
Expected Sample Size (Null Ref)	304.5
Expected Sample Size (Alt Ref)	230.0

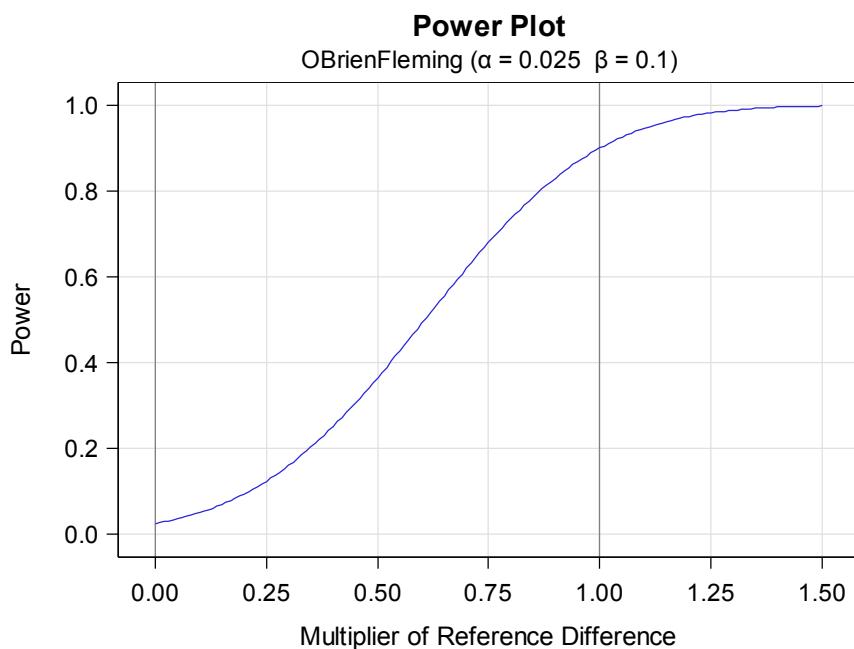
Any group sequential design is uniquely defined by a set of *stopping probabilities*, also known as *exit probabilities*. They are defined as the likelihood of crossing the stopping boundary (or boundaries) at each interim look under the null or alternative hypotheses. Table 6.6 provides information on the stopping probabilities in the group sequential design set up by Program 6.1. Examining the stopping probabilities under the null ( $CRef = 0$ ) helps us understand how this group sequential plan “spends” the overall Type I error rate  $\alpha$  throughout the course of a trial. Also, focusing on the alternative hypothesis ( $CRef = 1$ ), we see that, with the group sequential design based on the O'Brien-Fleming approach, early stopping is fairly unlikely. Specifically, under the alternative hypothesis, there is only a 29.3% chance that the trial will be terminated due to superior treatment effect at the first interim analysis (Stage 1). The probability of terminating the trial due to superior efficacy at the second interim analysis is equal to 69.6% (Stage 2).

**TABLE 6.6 Output from Program 6.1: Stopping probabilities**

CRef	Expected Stopping Stage	Source	Stopping Probabilities		
			Stage 1	Stage 2	Stage 3
0.00	2.987	Reject Null	0.0021	0.0105	0.0250
1.00	2.011	Reject Null	0.2928	0.6960	0.9000

It is also instructive to examine the power and average sample size of the resulting group sequential design. Figures 6.8 and 6.9 depict the power and average sample size of this design as a function of the true mean treatment difference. Note that this treatment difference is defined relative to the reference difference, i.e., the mean treatment difference assumed under the alternative hypothesis of no effect. This means that the value of 1 on the horizontal axis in either figure corresponds to the mean treatment difference of 3 points on the HAMD17 scale. Figure 6.8 displays the power function of the group sequential design, and it can be shown that it is virtually identical to that of the traditional designs with 150 patients per trial arm. Power of the O'Brien-Fleming design remains above 80% if the true mean treatment difference exceeds 2.7 points.

**Figure 6.8**  
**Case study 1**

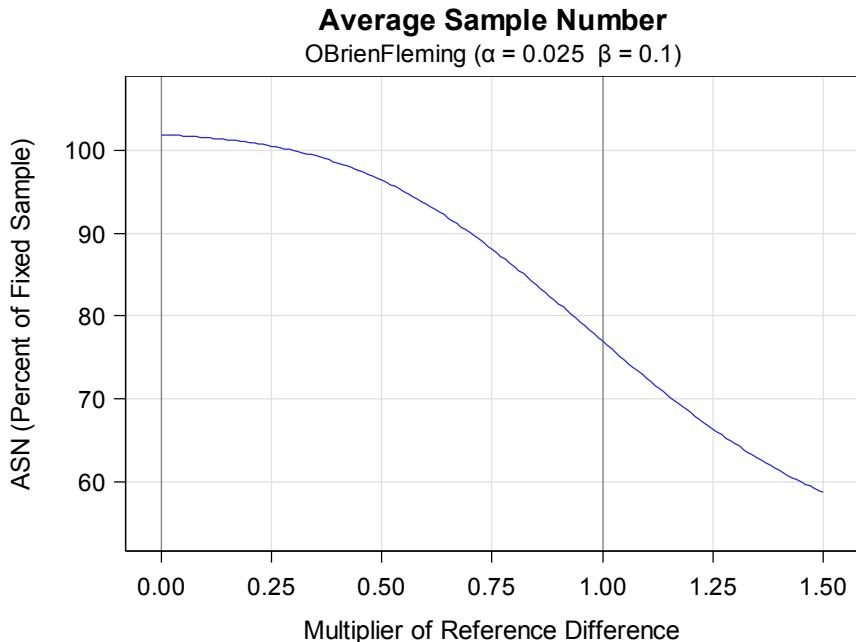


*Power of the O'Brien-Fleming design as a function of the true mean treatment difference. (The value of 1 on the horizontal axis corresponds to the mean treatment difference of 3 points.)*

Figure 6.9 indicates that the average sample size of the group sequential design is less than the sample size of 150 patients per arm that would have been used in the traditional design except for treatment effects in a small neighborhood of the null hypothesis, which corresponds to the value of 0 on the horizontal axis. Use of the O'Brien-Fleming plan results in appreciable savings in terms of the expected number of patients that will be enrolled in the trial if the treatment effect is large.

For example, in the extreme case when the true mean treatment difference is 4.5 points (which corresponds to the value of 1.5 on the horizontal axis), the average sample size in the group sequential design will be 42% less than the sample size used in the traditional design.

**Figure 6.9**  
Case study 1



*Average sample size of the O'Brien-Fleming design as a function of the true mean treatment difference (the value of 1 on the horizontal axis corresponds to the mean treatment difference of 3 points).*

It was pointed out above that Program 6.1 relied on a fixed data monitoring strategy to compute an O'Brien-Fleming stopping boundary. This was accomplished using `method=obf`. If `method` is set to `errfuncobf`, a flexible data monitoring strategy based on an O'Brien-Fleming-type spending function introduced in Lan and DeMets (1983) (see Section 6.2.4) will be used. This  $\alpha$ -spending function provides a good approximation to the original stopping boundaries introduced in O'Brien and Fleming (1979). Thus, the resulting group sequential design will be generally similar to that presented above.

#### PROGRAM 6.2 Group sequential design with an O'Brien-Fleming spending function in Case study 1

```
proc seqdesign stopprob(cref=0 1)
plots=(boundary(hscale=samplesize) power asn);
O'BrienFleming: design method=errfuncobf
nstages=3
info=cum(0.5 0.75 1)
alt=upper
stop=reject
alpha=0.025
beta=0.1;
samplesize model=twosamplemeans (meandiff=3 stddev=8);
run;
```

It is helpful to provide a quick comparison of key characteristics of the group sequential designs based on the fixed and flexible data monitoring strategies. To facilitate this comparison, Table 6.7 summarizes the general design information and provides a sample size summary. The information presented in this table matches that displayed in Table 6.5, which includes output from Program 6.1. The main characteristics of the group sequential design generated by Program 6.2, including the expected sample size under the null and alternative hypotheses, are virtually identical to those presented in Table 6.5.

**TABLE 6.7 Output from Program 6.2: Design information and Sample size summary**

Design information	
Statistic Distribution	Normal
Boundary Scale	Standardized Z
Alternative Hypothesis	Upper
Early Stop	Reject Null
Method	Error spending
Boundary Key	Both
Alternative Reference	3
Number of Stages	3
Alpha	0.025
Beta	0.1
Power	0.9
Max Information (Percent of Fixed Sample)	101.8
Max Information	1.189
Null Ref ASN (Percent of Fixed Sample)	101.5
Alt Ref ASN (Percent of Fixed Sample)	77.8

Sample size summary	
Test	Two-Sample Means
Mean Difference	3
Standard Deviation	8
Max Sample Size	304.3
Expected Sample Size (Null Ref)	303.5
Expected Sample Size (Alt Ref)	232.6

Further, Table 6.8 includes information on the stopping boundary derived from the Lan-DeMets  $\alpha$ -spending function. This stopping boundary is fairly similar to that computed using the fixed data monitoring strategy implemented in Program 6.1. It follows from Table 6.8 that the adjusted critical values at the three decision points computed from the Lan-DeMets  $\alpha$ -spending function are given by

$$u_1(0.025) = 2.963, \quad u_2(0.025) = 2.360, \quad u_3(0.025) = 2.014.$$

Comparing these boundary values to those displayed in Table 6.4, we see that the approach based on the  $\alpha$ -spending function results in a higher adjusted critical value at the first interim analysis. However, the difference between the fixed and flexible data monitoring strategies becomes quite trivial at the final analysis.

**TABLE 6.8 Output from Program 6.2: Boundary information**

Stage	Boundary Information (Standardized Z Scale)				
	Information Level			Alternative Reference	Boundary Values
	Proportion	Actual	N	Upper	Upper Alpha
1	0.50	0.594	152.2	2.313	2.963
2	0.75	0.892	228.3	2.833	2.360
3	1.00	1.189	304.3	3.271	2.014

### 6.2.6 Design stage: Pocock group approach in Case study 1

The O'Brien-Fleming design set up in Section 6.2.5 represents a rather conservative approach to conducting group sequential trials. The other extreme corresponds to the Pocock design. With a group sequential design based on a general Pocock plan, the Type I error rate is spent rapidly early in the trial. As a result, the trial is more likely to be terminated early due to superior efficacy at an interim analysis compared to an O'Brien-Fleming plan.

Program 6.3 uses PROC SEQDESIGN to construct a group sequential design with a fixed Pocock stopping boundary in Case study 1. All parameters of PROC SEQDESIGN in this program are identical to those used in Program 6.1 with the exception of the specification of the stopping boundary. A fixed data monitoring strategy with a Pocock stopping boundary is requested using `method=poc`. A more flexible approach based on a Pocock-type spending function can be applied by setting `method=errfuncpoc`. Also, as in Program 6.1, the resulting efficacy stopping boundary is saved in a data set (`poc_boundary`). This data set will play a key role in defining an efficacy monitoring strategy based on the group sequential design derived in Program 6.3 (see Section 6.2.8).

#### PROGRAM 6.3 Group sequential design with a fixed Pocock stopping boundary in Case study 1

```
proc seqdesign stopprob(cref=0 1)
    plots=(boundary(hscale=samplesize) power asn);
Pocock: design method=poc
    nstages=3
    info=cum(0.5 0.75 1)
    alt=upper
    stop=reject
    alpha=0.025
    beta=0.10;
samplesize model=twosamplemeans (meandiff=3 stddev=8);
ods output boundary=poc_boundary;
run;
```

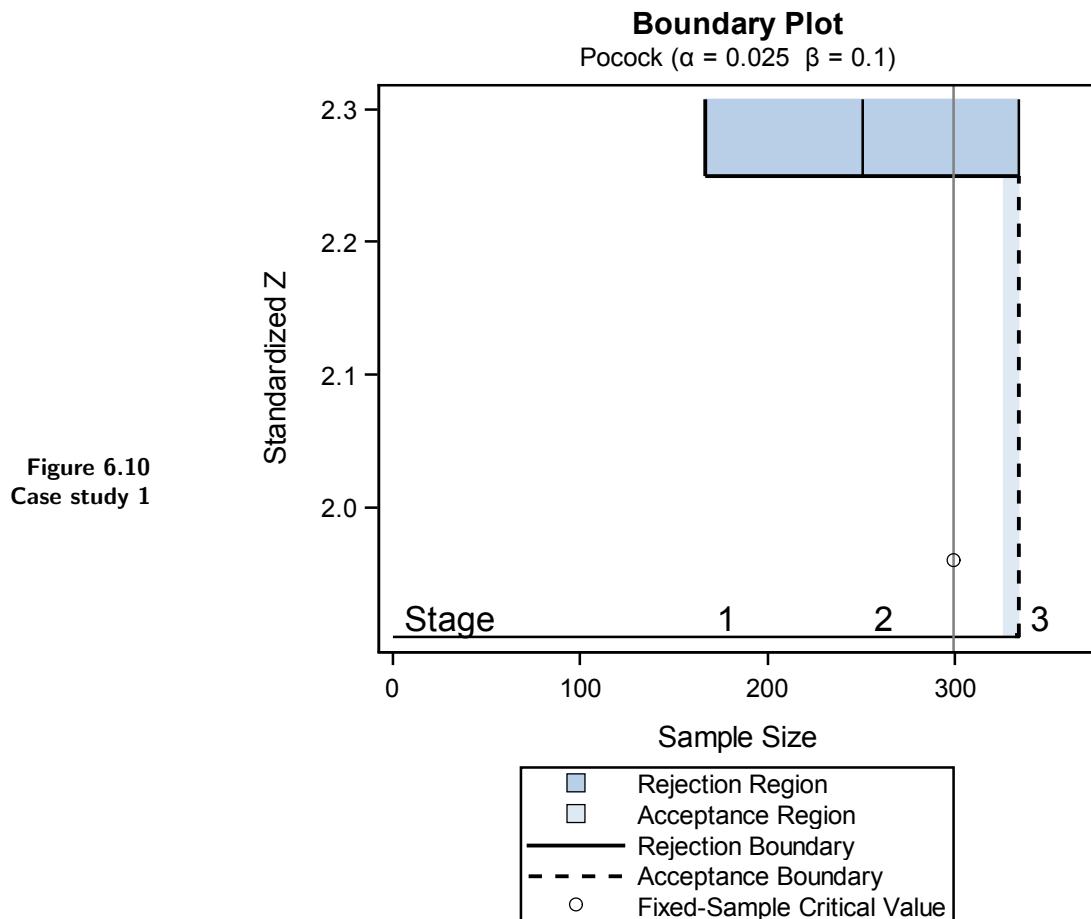
As in Section 6.2.5, the output generated by PROC SEQTEST is organized using three tables (Tables 6.9, 6.10, and 6.11). Beginning with Table 6.9, we see that a common adjusted critical value is used at all three decision points in the design based on the Pocock spending function. The critical values are given by

$$u_1(0.025) = 2.250, u_2(0.025) = 2.250, u_3(0.025) = 2.250.$$

Figure 6.10 displays the adjusted critical values (stopping boundaries) used in the Pocock group sequential design as a function of the total sample size. As explained above, the Pocock design relies on adjusted critical values that are constant across

**TABLE 6.9 Output from Program 6.3: Boundary information**

Stage	Boundary Information (Standardized Z Scale)				
	Information Level			Alternative Reference	Boundary Values
	Proportion	Actual	N		
1	0.50	0.652	167.0	2.42289	2.250
2	0.75	0.978	250.5	2.96742	2.250
3	1.00	1.305	334.0	3.42649	2.250



**Figure 6.10**  
**Case study 1**

*Stopping boundary of the Pocock design and critical value of the traditional design on a test statistic scale.*

the three decision points. An important implication of this property of the Pocock plan is that the Pocock-adjusted critical values are smaller than the O'Brien-Fleming-adjusted critical values early in the trial. This also implies that the Pocock test has less power to detect a significant difference at the final analysis. It is easy to check that, with the Pocock boundary, a one-sided  $p$ -value will be declared significant at the scheduled termination point if it is less than 0.0122. The corresponding  $p$ -value threshold at the final analysis in the O'Brien-Fleming design is substantially higher, i.e., 0.0215.

The key characteristics of the Pocock group sequential design are summarized in Table 6.10. A quick comparison with Table 6.5 reveals that the use of the Pocock stopping boundary results in an approximately 10% increase in the maximum sample size (from 305.5 to 334.0). The same is also true for the average sample size under the null hypothesis of no treatment effect. On the other hand, we can see from Tables 6.5 and 6.10 that, under the alternative hypothesis of a beneficial treatment effect (Alt Ref), the average sample sizes associated with the O'Brien-Fleming and Pocock plans are 230.0 and 220.8, respectively. This means that we should expect to enroll fewer patients in a group sequential trial with the Pocock stopping boundaries when the experimental treatment is truly efficacious.

To see why this happens, recall that the Pocock stopping boundary offers a higher

**TABLE 6.10 Output from Program 6.3: Design information and Sample size summary**

Design information		
Statistic Distribution		Normal
Boundary Scale		Standardized Z
Alternative Hypothesis		Upper
Early Stop		Reject Null
Method		Pocock
Boundary Key		Both
Alternative Reference		3
Number of Stages		3
Alpha	0.025	
Beta	0.1	
Power	0.9	
Max Information (Percent of Fixed Sample)	111.7	
Max Information	1.305	
Null Ref ASN (Percent of Fixed Sample)	110.9	
Alt Ref ASN (Percent of Fixed Sample)	73.9	
Sample size summary		
Test	Two-Sample Means	
Mean Difference	3	
Standard Deviation	8	
Max Sample Size	334.0	
Expected Sample Size (Null Ref)	331.3	
Expected Sample Size (Alt Ref)	220.8	

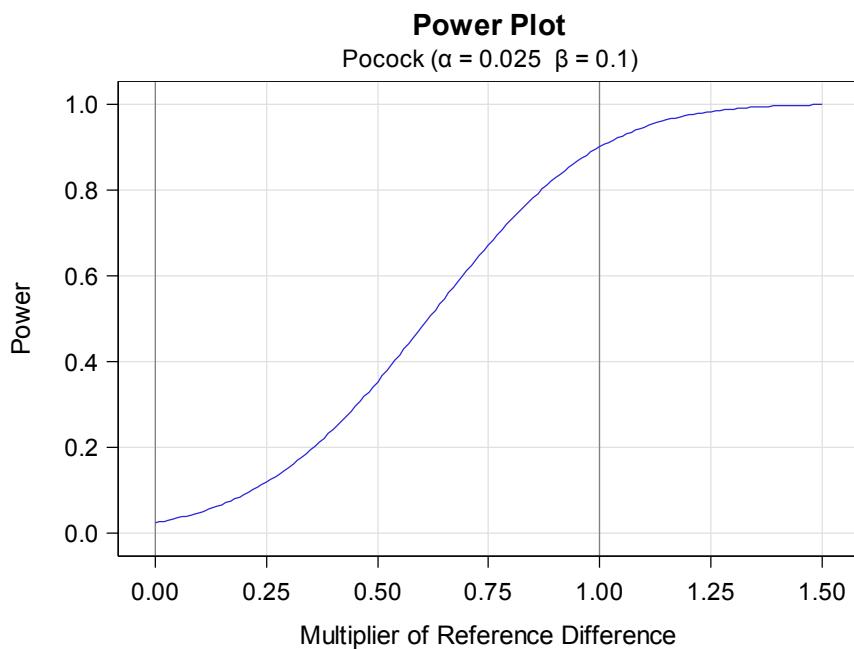
likelihood of early stopping at an interim analysis under the alternative hypothesis. Indeed, focusing on the probability of stopping due to superior efficacy under the alternative ( $CRef=1$ ) in Table 6.11, it is easy to see that the probability of stopping at the first interim analysis (Stage 1) is 56.9%. Thus, it is almost twice as high as the one associated with the O'Brien-Fleming stopping boundary (see Table 6.6). As a consequence, group sequential trials with a Pocock boundary are likely to be shorter in duration if the experimental treatment is associated with a strong efficacy profile.

**TABLE 6.11 Output from Program 6.3: Stopping probabilities**

CRef	Expected Stopping Stage	Source	Expected Cumulative Stopping Probabilities Reference = CRef * (Alt Reference)		
			Stage 1	Stage 2	Stage 3
0.00	2.968	Reject Null	0.0122	0.0194	0.0250
1.00	1.645	Reject Null	0.5687	0.7861	0.9000

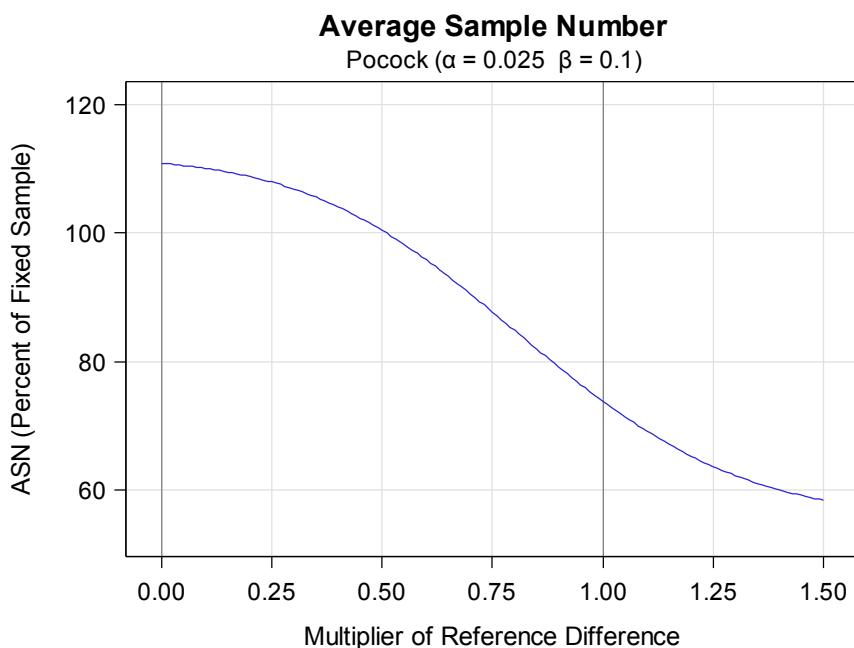
Figures 6.11 and 6.12 plot the power function and average sample size of the Pocock group sequential design derived in Program 6.3. It is easy to verify that the power function of the Pocock design is virtually identical to that of the O'Brien-Fleming design (see Figure 6.8). Further, considering Figure 6.12, recall that the Pocock design requires more patients when the treatment difference is equal to zero compared to the O'Brien-Fleming design and has a slight sample size advantage under the alternative hypothesis (i.e., when the mean treatment difference is 3 points). A comparison of Figures 6.9 and 6.12 reveals that the magnitude of this advantage is quite small, and we can argue that the Pocock design is generally inferior to the O'Brien-Fleming design in terms of the average sample size when the true treatment effect is less than its hypothesized value.

**Figure 6.11**  
Case study 1



*Power of the Pocock design as a function of the true mean treatment difference (the value of 1 on the horizontal axis corresponds to the mean treatment difference of 3 points).*

**Figure 6.12**  
Case study 1



*Average sample size of the Pocock design as a function of the true mean treatment difference (the value of 1 on the horizontal axis corresponds to the mean treatment difference of 3 points).*

## Pocock and O'Brien-Fleming group sequential designs

After we have reviewed group sequential designs with Pocock and O'Brien-Fleming stopping boundaries, it is natural to ask for recommendations or rules of thumb on how to pick a set of stopping boundaries for a particular clinical trial. There is really no hard and fast rule, and the answer to this question depends heavily on the objectives of each individual trial. The following is a set of general guidelines that can help us select an appropriate set of stopping boundaries in most clinical trials.

- **Clinical considerations.** The magnitude of the treatment effect observed in a clinical trial needs to be balanced against the size of the sample it was estimated from. Only an absolutely overwhelming treatment difference can justify the termination of a clinical trial after a quarter or a third of the patients have been enrolled. Additionally, as was pointed out in the Introduction, early stopping of a trial complicates a full characterization of the safety profile of the treatment. A decision to terminate a trial prematurely is clearly undesirable when one wishes to examine the effect of an experimental therapy on multiple safety and secondary efficacy variables in an adequately-sized study. To stress this point, Emerson and Fleming (1989) noted that

A design that suggests early termination only in the presence of extreme evidence for one treatment over another will allow greatest flexibility to examine other response variables or long-term effects while maintaining adequate treatment of the ethical concerns in clinical trials.

Along the same line, O'Brien (1990) indicated that

One of the disadvantages of early termination is that it may preclude obtaining satisfactory answers to secondary (but nonetheless important) questions pertaining to drug safety and efficacy.

In view of these considerations, it should not be surprising that stopping boundaries are typically chosen to minimize the chances of early trial termination. Thus, group sequential plans with O'Brien-Fleming boundaries are used more frequently in clinical trials than plans with Pocock boundaries.

- **Sample size considerations.** Another important consideration in a group sequential setting is the average and maximum number of patients that will be enrolled into the trial. It is known that sequential plans based on Pocock boundaries have a smaller average sample size when the true treatment difference is large. Using O'Brien-Fleming boundaries results in a smaller maximum sample size and smaller average sample size under the null hypothesis of no treatment difference. A smaller expected sample size is easily achieved by increasing the probability of early termination. For example, Wang and Tsiatis (1987) studied group sequential tests that minimize the expected sample size under the alternative hypothesis. Wang and Tsiatis concluded that the Pocock test is nearly optimal in most practical situations (e.g., in clinical trials powered at 80% and 90%). However, in general, mathematical considerations, such as optimization of the average sample size in a group sequential procedure, are less important than the clinical considerations outlined above.
- **Data management considerations.** Quite often, the first interim analysis is conducted to test the data management process, check compliance and protocol violations, and ensure the quality of the data received by a data monitoring committee. Once it has been verified that the interim data are reliable, a data monitoring committee will have more confidence in the data reported at subsequent

interim analyses. In order to avoid early stopping for efficacy at the first interim analysis, we can use the O'Brien-Fleming or similar boundary. Later interim analyses can be designed to detect early evidence of efficacy, in which case the O'Brien-Fleming boundary is again preferable to the Pocock boundary because it maximizes the probability of stopping in the second half of the study if the experimental treatment is truly efficacious.

There might be other considerations that determine the choice of a sequential testing plan. For example, in clinical trials with a futility stopping boundary, the trial's sponsor would generally hesitate to go for an "early kill" of a treatment with a new mechanism of action since an "early kill" will have negative implications for all drugs in the same class. Because of this, the sponsor will most likely vote for an O'Brien-Fleming-type stopping boundary. On the other hand, it takes less evidence in order to convince clinical researchers to stop developing a treatment with a well-understood mechanism of action.

### **6.2.7 Design stage: Efficacy and futility testing in Case study 2**

Thus far, we have considered group sequential designs for detecting early signs of superior efficacy. It is also helpful to quickly review group sequential designs that support an option to terminate the trial before its planned end due to apparent futility of the experimental therapy. As pointed out in Section 6.2.1, it is fairly uncommon to use a repeated significance testing approach in Phase III trials with decision rules based on futility and efficacy assessments because this leads to binding futility stopping rules that cannot be overridden. Most trial sponsors prefer to set up futility stopping rules in a non-binding manner, which means that the sponsor reserves the right to continue the trial even if the predefined futility stopping rule is met. To set up group sequential designs with non-binding futility stopping rules, a repeated significance testing approach is applied to define a stopping boundary for efficacy assessment that preserves the Type I error rate. After this, a futility stopping rule is set up at each interim analysis independently of the efficacy stopping boundary. This results in a set of decision rules with a deflated Type I error rate and an inflated Type II error rate.

To illustrate the process of constructing group sequential plans with efficacy and futility stopping boundaries with prespecified Type I and Type II error rates, the Phase III trial in patients with severe sepsis introduced in Section 6.2.2 will be employed. Three decision points were considered in this trial (two interim looks and final analysis). The first analysis was intended to be used primarily for futility testing. The trial was to be stopped due to superior efficacy only in the presence of an overwhelming treatment difference. The second analysis was to involve efficacy and futility assessments.

A group sequential plan that meets both of these objectives can be set up by using an O'Brien-Fleming stopping boundary for testing the efficacy of the experimental treatment and a Pocock boundary for futility analyses. As we showed in Section 6.2.5, it is very difficult to exceed O'Brien-Fleming critical values early in the trial. Thus, using the O'Brien-Fleming boundary will minimize the chances of establishing superior efficacy at the first interim inspection. Furthermore, this boundary will improve the power of treatment comparisons at the second interim look and the final analysis. On the other hand, using a Pocock stopping boundary for futility assessment ensures that the trial can be terminated early if the observed treatment difference is negative or too small. The described group sequential design is consistent with the principle of futility testing outlined by DeMets and Ware (1980):

When early evidence suggests the new therapy may be inferior, study termination and acceptance of [the null hypothesis] may be indicated before the statistic reaches a value as extreme as that required to reject [the null hypothesis] in favour of [the alternative hypothesis]. In other words, we are willing to terminate and accept [the null hypothesis] with less evidence that treatment could be harmful than we would require to claim a benefit.

Program 6.4 computes the key characteristics of the described sequential design using PROC SEQDESIGN. As in Programs 6.1 and 6.3, the main components of the design need to be defined in this program. First, the following statement

```
method(alpha)=obf method(beta)=poc
```

indicates that the efficacy stopping boundary will be derived using the O'Brien-Fleming approach, and the Pocock approach will be used to set up the futility stopping boundary. The number of decision points is specified using `nstages=3`. Since the two interim analyses are planned to occur after 20% and 66% of the patients completed the 28-day study period, the information fractions at the three decision points are set to `info=cum(0.2 0.66 1)`. The upper stopping boundary will be used for efficacy assessment, which means that `alt=upper` and `stop=both` indicate that the trial will be terminated if either boundary is crossed. Finally, the one-sided Type I error rates in the severe sepsis trial are equal to 0.025, and the trial is powered at 80%. This means that `alpha=0.025` and `beta=0.2`.

#### **PROGRAM 6.4 Group sequential plan for efficacy and futility testing in Case study 2**

```
proc seqdesign stopprob(cref=0 1)
plots=(boundary(hscale=samplesize) power asn);
OBF_Pocock: design method(alpha)=obf method(beta)=poc
nstages=3
info=cum(0.2 0.66 1)
alt=upper
stop=both
alpha=0.025
beta=0.20;
samplesize model=twosamplefreq (test=prop
nullprop=0.7 prop=0.76
ref=avgprop);
ods output boundary=obf_poc_boundary;
run;
```

Power calculations will be performed by PROC SEQDESIGN using the treatment effect assumptions defined in Section 6.2.2. Switching to 28-day survival rates, the assumed survival rates in the placebo and treatment arms are 70% and 76%, respectively. The survival rates in the two arms will be compared using a two-sample test for proportions. Thus, power calculations will be run based on the following specifications:

```
samplesize model=twosamplefreq
(test=prop nullprop=0.7 prop=0.76 ref=avgprop);
```

Note that `ref=avgprop` specifies that the average alternative proportion should be used in the calculations.

Table 6.12 displays the main characteristics of the group sequential design in the severe sepsis trial. First, let us examine the effect of the futility boundary on the expected sample size under the null and alternative hypotheses. Recall from

Table 6.5 that group sequential designs with an O'Brien-Fleming boundary require on average fewer patients than traditional designs with a fixed sample size when the experimental treatment is truly efficacious. This advantage, however, is lost under the null hypothesis. Table 6.12 demonstrates that adding a lower stopping boundary completely changes the behavior of O'Brien-Fleming group sequential plans. Because of the possibility of early termination due to futility, the average sample size of the group sequential design under the null hypothesis of no treatment effect (Null Ref) is 928.1 patients. This sample size is substantially smaller than the total sample size in the traditional derived in Section 6.2.2, i.e., 1718 patients. The group sequential design is also superior to the traditional design under the alternative hypothesis (Alt Ref). The only downside of simultaneous efficacy/futility testing strategies is an increased maximum sample size. We see from Table 6.12 that the sequential plan can potentially require 25% more patients than the traditional design (2155.7 patients versus 1718 patients).

**TABLE 6.12 Output from Program 6.4: Design information and Sample size summary**

Design information	
Statistic Distribution	Normal
Boundary Scale	Standardized Z
Alternative Hypothesis	Upper
Early Stop	Accept/Reject Null
Method	Unified Family
Boundary Key	Both
Alternative Reference	0.06
Number of Stages	3
Alpha	0.025
Beta	0.2
Power	0.8
Max Information (Percent of Fixed Sample)	125.4
Max Information	2734.3
Null Ref ASN (Percent of Fixed Sample)	54.0
Alt Ref ASN (Percent of Fixed Sample)	87.6

Sample size summary	
Test	Two-Sample Proportions
Null Proportion	0.7
Proportion (Group A)	0.76
Test	Statistic Z for Proportion
Reference Proportions	Alt Ref (Average Proportion)
Max Sample Size	2155.7
Expected Sample Size (Null Ref)	928.1
Expected Sample Size (Alt Ref)	1505.5

Table 6.13 presents the lower and upper adjusted critical values (Boundary values) at the two interim looks and final analysis. The lower and upper critical values are shown in the table in the columns labeled “Beta” and “Alpha”, respectively. As expected, the upper critical value at the first interim inspection is extremely large ( $u_1 = 4.1765$ ) and will be exceeded only if the treatment difference is absolutely phenomenal. The corresponding lower critical value is positive ( $l_1 = 0.1335$ ). Thus, the trial will be terminated at the first look due to futility unless the experimental treatment demonstrates some degree of efficacy. The futility rule is well justified in this severe sepsis trial since it is ethically unacceptable to expose patients with severe conditions to ineffective therapies. Moving on to the second interim look, we see that the continuation interval has become much narrower. The experimental treatment will be declared futile if it does not meet minimum efficacy requirements ( $l_2 = 1.2792$ ) or superior to placebo if the test statistic is greater than  $u_2 = 2.2991$ . Finally, the trial will end with a positive outcome at the final analysis if the test

**TABLE 6.13 Output from Program 6.4: Boundary information**

Stage	Boundary Information (Standardized Z Scale)					
	Information Level			Alternative Reference	Boundary Values	
	Proportion	Actual	N		Upper	Beta
1	0.20	546.9	431.1	1.4031	0.1335	4.1765
2	0.66	1804.6	1422.8	2.5489	1.2792	2.2991
3	1.00	2734.3	2155.7	3.1374	1.8678	1.8678

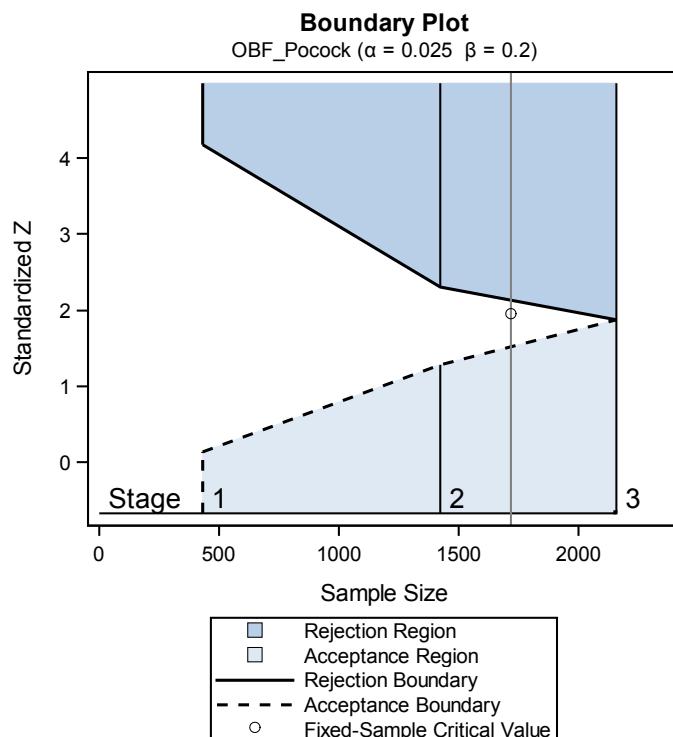
statistic exceeds  $u_3 = 1.8678$ , or, equivalently, the final  $p$ -value is less than 0.0309.

Note that the  $p$ -value cutoff used at the final analysis is greater than the familiar 0.025 cutoff that would have been used in a traditional design. This happens because the binding futility stopping rule results in a deflated Type I error probability, and the critical value at the final analysis needs to be adjusted upward to keep the one-sided Type I error rate at the 0.025 level. See Chang and Chuang-Stein (2004) for a detailed discussion of this phenomenon.

It is important to note that, with a binding futility stopping rule, the trial's sponsor no longer has an option to override the rule. If the futility boundary is crossed and the trial sponsor chooses to continue the trial, therapeutic benefit cannot be claimed under any circumstances. The upper and lower stopping boundaries were created in this setting under the assumption that crossing the lower stopping boundary would require termination. See Lan, Lachin, and Bautista (2003) for a further discussion of this problem.

A graphical summary of the stopping boundaries on a test statistic scale is presented in Figure 6.13. The figure shows that the continuation region, i.e., the

**Figure 6.13**  
Case study 1



*Efficacy and futility stopping boundaries of the group sequential design and critical value of the traditional design on a test statistic scale.*

interval between the upper and lower adjusted critical values, is quite wide at the first interim look. This indicates that the trial will be terminated at this look only if the observed efficacy signal is very strong or fairly weak. The continuation region shrinks to a rather narrow interval at the second interim analysis, and the two stopping boundaries converge at the final analysis.

The stopping probabilities at the three decision points are listed in Table 6.14. As before, the probabilities evaluated under the null hypothesis of no treatment benefit correspond to CRef=0. Assuming that the experimental treatment does not improve patients' survival, it follows from this table that there is a 55.3% chance that a decision to stop the trial due to futility will be made at the very first interim look (Stage 1). The probability of early stopping due to lack of efficacy increases to 91.7% by the second interim look (Stage 2), which virtually guarantees that the trial will be terminated if the treatment is truly ineffective. When the final analysis is considered (Stage 3), the probability of rejecting the null hypothesis is simply equal to the Type I error rate in the trial.

Considering the case when the alternative hypothesis is true (CRef=1), Table 6.14 shows that, as expected, there is an extremely small chance, namely, 0.2%, of claiming a statistically significant effect at the first interim analysis. The probability of stopping the trial due to overwhelming efficacy is, in fact, much lower than the probability of stopping due to futility (10.2%). Since the continuation region is quite narrow at the second look, the trial is quite likely to be stopped with the conclusion that the experimental treatment is effective. The probability of this outcome is 57.8%. By contrast, the probability of early stopping due to lack of efficacy grows much more slowly and reaches 16.8% at the second interim look. Lastly, the probability of accepting the null hypothesis at the final analysis is the Type II error rate of the group sequential design.

**TABLE 6.14 Output from Program 6.4: Stopping probabilities**

CRef	Expected Stopping Stage	Source	Expected Cumulative Stopping Probabilities				
			Reference = CRef * (Alt Reference)	Stopping Probabilities	Stage 1	Stage 2	Stage 3
0.00	1.501	Reject Null	0.00001	0.0101	0.0250		
0.00	1.501	Accept Null	0.5531	0.9165	0.9750		
0.00	1.501	Total	0.5531	0.9266	1.0000		
1.00	2.113	Reject Null	0.0028	0.5776	0.8000		
1.00	2.113	Accept Null	0.1021	0.1676	0.2000		
1.00	2.113	Total	0.1049	0.7452	1.0000		

### 6.2.8 Monitoring stage: O'Brien-Fleming and Pocock spending functions in Case study 1

The error spending approach was introduced in Section 6.2.4 in the context of designing group sequential trials. The same methodology can be applied at the monitoring stage of a trial. It was emphasized in Section 6.2.4 that, by construction, the error spending approach supports flexible interim monitoring strategies where the timing of interim looks does not need to be prespecified at the design stage.

However, it is important to remember that the theoretical properties of the error spending approach hold under the assumption that the timing of future looks is independent of what has been observed in the past. In theory, the overall probability of Type I errors might no longer be preserved if we modify the sequential testing scheme due to promising findings at one of the interim looks. Several authors have studied the overall Type I error rate of sequential plans in which this assumption

is violated. For example, Lan and DeMets (1989) described a realistic scenario in which a large but nonsignificant test statistic at an interim look causes the data monitoring committee to request additional looks at the data. Lan and DeMets showed via simulations that data-dependent changes in the frequency of interim analyses generally have a minimal effect on the overall  $\alpha$  level. Further, as pointed out by Proschan, Follman, and Waclawiw (1992), the error spending approach provides a certain degree of protection against very frequent looks at the data. The amount of Type I error rate that can be spent at each analysis is roughly proportional to the amount of information accrued since the last look. Even though the test statistic is close to the stopping boundary at the current look, there is a good chance it will not cross the boundary at the next look if that is taken immediately after the current one.

Data monitoring in a group sequential setting based on the error spending approach is illustrated below using the depression trial from Case study 1 (see Section 6.2.8). Efficacy monitoring methods will be implemented using PROC SEQTEST. Related important topics such as the computation of confidence intervals for the treatment difference at each interim analysis as well as derivation of the bias-adjusted point estimate of the treatment difference at the very last analysis in the trial will be described later in Sections 6.2.9 and 6.2.10.

## Interim monitoring based on the O'Brien-Fleming approach in Case study 1

The implementation of group sequential monitoring strategies based on a pre-defined  $\alpha$ -spending function will use the designs based on the popular spending functions (O'Brien-Fleming and Pocock spending functions) that were introduced in Sections 6.2.5 and 6.2.6.

To set the stage for the discussion of interim monitoring in the depression trial, recall that stopping boundaries can be defined on either a test statistic or  $p$ -value scale. Thus, it is helpful to begin by computing a normally distributed statistic and associated  $p$ -value at each decision point (interim analysis or final analysis) in the trial.

Consider the following  $z$  statistic for testing the null hypothesis of no difference between the treatment and placebo at one of the three decision points in the depression trial

$$Z_k = \frac{(\bar{X}_{1k} - \bar{X}_{2k})}{s_k \sqrt{1/n_{1k} + 1/n_{2k}}}.$$

Here  $n_{1k}$  and  $\bar{X}_{1k}$  ( $n_{2k}$  and  $\bar{X}_{2k}$ ) denote the number of patients and sample mean in the experimental (placebo) arm at the  $k$ th decision point ( $k = 1, 2, 3$ ). Also,  $s_k$  is the pooled standard deviation, i.e.,

$$s_k = \sqrt{\frac{(n_{1k} - 1)s_{1k}^2 + (n_{2k} - 1)s_{2k}^2}{n_{1k} + n_{2k} - 2}}$$

with  $s_{1k}$  and  $s_{2k}$  denoting the standard deviations of HAMD17 changes from baseline in the experimental and placebo arms, respectively, at the  $k$ th decision point. Since  $Z_k$  is normally distributed, a one-sided  $p$ -value for testing the superiority of the experimental treatment to placebo at the  $k$ th look is given by

$$p_k = 1 - \Phi(Z_k).$$

It is important to note that we are by no means restricted to statistics from the two-sample  $z$ -test. The decision-making process described below can be used with

any normally or nearly normally distributed test statistic, e.g., a test statistic from an ANOVA model with multiple terms representing important prognostic factors.

Program 6.5 uses the formulas displayed above to compute the  $z$  statistics from the mean HAMD17 changes and associated standard deviations (see Table 6.1) at the three decision points. In addition, this program uses information from the group sequential design based on the O'Brien-Fleming approach derived in Section 6.2.5. Specifically, the maximum amount of information (1.193) and maximum sample size in the group sequential trial are taken from Table 6.5.

### **PROGRAM 6.5 Test statistics and associated $p$ -values in Case study 1**

```

data DepTrialData;
  input _Stage_ n1 mean1 sd1 n2 mean2 sd2;
  datalines;
  1 78 8.3 6.2 78 5.9 6.5
  2 122 8.0 6.3 120 6.3 5.9
  3 150 8.2 5.9 152 6.1 5.8
run;

data obf_teststat;
  set DepTrialData;
  _Scale_= 'StdZ';
  maxinfo=1.193;
  maxn=305.5;
  _info_= maxinfo* (n1+n2) /maxn;
  s= sqrt(((n1-1)*sd1*sd1+(n2-1)*sd2*sd2)/(n1+n2-2));
  StdErr= s * sqrt(1/n1+1/n2);
  ZStat= (mean1-mean2) / StdErr;
  keep _Scale_ _Stage_ _Info_ ZStat;
run;

proc print data=obf_teststat;
run;

```

---

<b>Output from Program 6.5</b>	<table border="1"> <thead> <tr> <th>_Stage_</th><th>_Scale_</th><th>_info_</th><th>ZStat</th></tr> </thead> <tbody> <tr> <td>1</td><td>StdZ</td><td>0.6092</td><td>2.3597</td></tr> <tr> <td>2</td><td>StdZ</td><td>0.9453</td><td>2.1659</td></tr> <tr> <td>3</td><td>StdZ</td><td>1.1797</td><td>3.1192</td></tr> </tbody> </table>	_Stage_	_Scale_	_info_	ZStat	1	StdZ	0.6092	2.3597	2	StdZ	0.9453	2.1659	3	StdZ	1.1797	3.1192
_Stage_	_Scale_	_info_	ZStat														
1	StdZ	0.6092	2.3597														
2	StdZ	0.9453	2.1659														
3	StdZ	1.1797	3.1192														

---

Output 6.5 displays the information fraction (`_info_`) and test statistic (`ZStat`) at each decision point in the trial. It follows from the output that the test statistics are all highly significant at a conventional one-sided 0.025 level. However, it remains to be seen whether or not these statistics will stay significant after a proper adjustment for multiple analyses. In what follows, we will examine adjustments based on the O'Brien-Fleming and Pocock group sequential plans.

Program 6.6 performs group sequential monitoring of the depression trial based on the O'Brien-Fleming approach using PROC SEQTEST. The program creates the data set with the efficacy stopping boundary (`obf_boundary`) produced by Program 6.1 as well as the data set with the information fractions and treatment effect test statistics at the three decision points (`obf_teststat`) computed by Program 6.5. The three blocks of code in Program 6.6 perform efficacy assessments at the first interim analysis (`obs=1`), second interim analysis (`obs=2`), and final analysis (`obs=3`). Note that the `_Scale_` parameter was set to `StdZ` in the `obf_teststat` data set, which means that the stopping boundaries will be computed on a test statistic scale rather than a  $p$ -value scale.

**PROGRAM 6.6 Efficacy monitoring based on the O'Brien-Fleming approach in Case study 1**

```

proc seqtest boundary=obf_boundary
  data(testvar=ZStat)=obf_teststat(obs=1)
  nstages=3
  order=stagewise;
run;

proc seqtest boundary=obf_boundary
  data(testvar=ZStat)=obf_teststat(obs=2)
  nstages=3
  order=stagewise;
run;

proc seqtest boundary=obf_boundary
  data(testvar=ZStat)=obf_teststat(obs=3)
  nstages=3
  order=stagewise;
run;

```

Table 6.15 lists the adjusted critical value (boundary value) computed from the O'Brien-Fleming spending function as well as the observed treatment effect test statistic at the first interim look. The test statistic (2.3597) clearly lies below the adjusted critical value (2.8133), which indicates that the O'Brien-Fleming stopping boundary has not been crossed. As a result, the trial will continue to the next interim look.

**TABLE 6.15 Output from Program 6.6: First interim analysis**

Stage	Test Information (Standardized Z Scale)					
	Information Level		Alternative Reference	Boundary Values	Test	
	Proportion	Actual	Upper	Upper Alpha	Estimate	Action
1	0.5105	0.6092	2.3415	2.8133	2.3597	Continue
2	0.7553	0.9012	2.8480	2.3310	.	.
3	1.0000	1.1933	3.2772	2.0270	.	.

Table 6.16 displays the adjusted critical value (boundary value) at the second interim look. It is worth noting that this critical value has been updated compared to Table 6.15 to account for the actual timing of the interim analysis, i.e., the actual information level at this decision point. The revised stopping boundary on a test statistic scale at the second look is 2.2565. Again, even though the test statistic at the second look, which is equal to 2.1659, is clearly significant at a one-sided 0.025 level, it is not large enough to cross the stopping boundary. More evidence is needed before the experimental treatment can be declared superior to placebo with respect to its effect on the HAMD17 total score, and the trial will be continued to the final analysis.

**TABLE 6.16 Output from Program 6.6: Second interim analysis**

Stage	Test Information (Standardized Z Scale)					
	Information Level		Alternative Reference	Boundary Values	Test	
	Proportion	Actual	Upper	Upper Alpha	Estimate	Action
1	0.5105	0.6092	2.3415	2.8133	2.3597	Continue
2	0.7919	0.9450	2.9164	2.2565	2.1659	Continue
3	1.0000	1.1933	3.2772	2.0416	.	.

Table 6.17 lists the adjusted critical value (boundary value) at the final analysis, which was slightly modified compared to Table 6.16 to take into account the fact that the actual information level at this analysis is slightly different from the projected information level. The updated stopping boundary at the final analysis shown in Table 6.17 is 2.0380, and the test statistic (3.1192) clearly exceeds this stopping boundary. Therefore, we conclude that the mean treatment difference in HAMD17 changes is significantly different from zero at a one-sided 0.025 level.

**TABLE 6.17 Output from Program 6.6: Final analysis**

Stage	Test Information (Standardized Z Scale)					
	Information Level		Alternative Reference	Boundary Values	Test	
	Proportion	Actual	Upper	Upper	ZStat	Action
1	0.5105	0.6092	2.3415	2.8133	2.3597	Continue
2	0.7919	0.9450	2.9164	2.2565	2.1659	Continue
3	1.0000	1.1793	3.2579	2.0380	3.1192	Reject Null

As was noted before, one of the attractive features of O'Brien-Fleming designs is that they “spend” very little Type I error probability early in the trial. Thus, the final analysis is performed at a significance level that is only marginally smaller than the corresponding unadjusted level used in traditional designs. Indeed, we can see from Table 6.17 that the adjusted critical value of 2.0380 is larger than the conventionally used value of 1.959. With the Pocock design, we should normally expect a much larger difference between the adjusted and unadjusted critical value.

### Interim monitoring based on the Pocock approach in Case study 1

It will be instructive to compare the O'Brien-Fleming monitoring strategy we considered above to a strategy based on the Pocock design presented in Section 6.2.6. As before, the first step in implementing efficacy monitoring in this clinical trial is to create a data set with information fractions and test statistics at the individual decision points that will be passed to PROC SEQTEST. This data set (*poc\_teststat*) was created using Program 6.7. The code in Program 6.7 is virtually the same as the code shown in Program 6.5. The only difference is that the maximum amount of information and maximum sample size in the group sequential trial were updated. The values used in Program 6.7, i.e., 1.305 and 334.0, came from Table 6.10 that summarizes key characteristics of the Pocock group sequential design.

#### PROGRAM 6.7 Efficacy monitoring based on the Pocock approach in Case study 1

```

data DepTrialData;
  input _Stage_ n1 mean1 sd1 n2 mean2 sd2;
  datalines;
  1 78 8.3 6.2 78 5.9 6.5
  2 122 8.0 6.3 120 6.3 5.9
  3 150 8.2 5.9 152 6.1 5.8
run;

data poc_teststat;
  set DepTrialData;
  _Scale_= 'StdZ';
  maxinfo=1.305;
  maxn=334.0;

```

```

_info_= maxinfo* (n1+n2) /maxn;
s= sqrt(((n1-1)*sd1*sd1+(n2-1)*sd2*sd2)/(n1+n2-2));
StdErr= s * sqrt(1/n1+1/n2);
ZStat= (mean1-mean2) / StdErr;
keep _Scale_ _Stage_ _Info_ ZStat;
run;

proc print data=poc_teststat;
run;

```

**Output from Program 6.7**

_Stage_	_Scale_	_info_	ZStat
1	StdZ	0.6092	2.3597
2	StdZ	0.9455	2.1659
3	StdZ	1.1800	3.1192

Output 6.7 lists the information fractions and test statistics in the trial. Note that the test statistics shown in this output are identical to those displayed in Output 6.5, but the information fractions have been slightly adjusted to account for the fact that the Pocock spending function was used to set up this group sequential design.

PROC SEQTEST is used in Program 6.8 to perform efficacy monitoring of the depression trial based on the Pocock group sequential design. As in Program 6.6, information on the original trial design and test statistics is passed to PROC SEQTEST using two data sets that were created in Programs 6.3 and 6.7 (i.e., poc\_boundary and poc\_teststat). The treatment effect assessment is performed in Program 6.8 only at the first interim analysis (obs=1).

**PROGRAM 6.8 Efficacy monitoring based on the Pocock approach in Case study 1**

```

proc seqtest boundary=poc_boundary
  data(testvar=ZStat)=poc_teststat(obs=1)
  nstages=3
  order=stagewise;
run;

```

Table 6.18 shows the adjusted critical value (boundary value) of the Pocock design along with the treatment effect test statistic at the first interim look. The table demonstrates that the treatment difference turns out to be significant at the very first look after only 156 patients completed the trial. The test statistic at the first interim analysis is 2.3597 and exceeds the monitoring boundary in this group sequential design (2.2757). The observed inconsistency between the O'Brien-Fleming and Pocock decision functions is not completely unexpected. Pocock stopping boundaries are known to be anti-conservative at the early interim looks and are likely to lead to a rejection of the null hypothesis, which is exactly what happened in this particular example.

**TABLE 6.18 Output from Program 6.8: First interim analysis**

Stage	Test Information (Standardized Z Scale)					
			Null Reference = 0			
	Information Level	Alternative Reference	Boundary Values	Test		
	Proportion	Actual	Upper	Upper	Alpha	Estimate
1	0.4672	0.6095	2.3422	2.2757	2.3597	Reject Null
2	0.7336	0.9570	2.9348	2.2561	.	.
3	1.0000	1.3045	3.4265	2.2418	.	.

### 6.2.9 Repeated confidence intervals

It has been pointed out before that an analysis of clinical trial data in a sequential manner is likely to greatly increase the overall Type I error rate. Thus, we need to adjust the critical values upward at each interim look. For the same reason, we should avoid using naive (unadjusted) confidence intervals in group sequential trials. Mathematically, unadjusted confidence intervals are too narrow to achieve the nominal coverage probability at multiple interim analyses. In order to alleviate this problem, Jennison and Turnbull (1989) proposed a simple method for adjusting the width of confidence intervals to increase the joint coverage probability under any sequential monitoring scheme. The resulting confidence intervals are known as *repeated confidence intervals*.

To define repeated confidence intervals, consider a group sequential trial with  $m$  looks. Suppose that a study is conducted to test the two-sided hypothesis that the true treatment difference  $\delta$  is equal to 0 (the treatment difference can be defined as a difference in means or a difference in proportions). Let  $\hat{\delta}_k$  and  $s_k$  denote the estimate of the treatment difference and its standard error at the  $k$ th look, respectively. In the repeated significance testing framework, the test statistic

$$Z_k = \hat{\delta}_k / s_k$$

is compared to the lower and upper adjusted critical values  $l_k$  and  $u_k$ , which are chosen in such a way that the probability of crossing the stopping boundaries does not exceed  $\alpha$  under the null hypothesis of no treatment effect:

$$P\{Z_k < l_k \text{ or } Z_k > u_k \text{ for any } k = 1, \dots, m \text{ if } \delta = 0\} = \alpha.$$

Jennison and Turnbull (1989) demonstrated that it is easy to invert this testing procedure in order to compute confidence intervals for the unknown treatment difference at each of the interim analyses. The two-sided repeated confidence intervals for  $\delta$  are given by

$$CI_k = (\hat{\delta}_k - u_k s_k, \hat{\delta}_k + l_k s_k), \quad k = 1, \dots, m,$$

and possess the following important property<sup>4</sup>. By the definition of the adjusted critical values, the *joint* coverage probability of the repeated confidence intervals is greater than or equal to  $1 - \alpha$ , i.e.,

$$P\{\hat{\delta}_k - u_k s_k \leq \delta \leq \hat{\delta}_k + l_k s_k \text{ for all } k = 1, \dots, m\} \geq 1 - \alpha.$$

regardless of the choice of stopping boundaries or any other aspects of a sequential monitoring scheme. As pointed out by Jennison and Turnbull (2000, Chapter 9), this immediately implies that the constructed confidence intervals are valid under any group sequential plan. Indeed, for any random stopping time  $\tau$ , the coverage probability is always maintained at the same level:

$$P\{\hat{\delta}_\tau - u_\tau s_\tau \leq \delta \leq \hat{\delta}_\tau + l_\tau s_\tau\} \geq 1 - \alpha.$$

Repeated confidence intervals in group sequential trials with a single upper boundary are defined in a similar manner. These intervals are set up as one-sided intervals and are chosen to satisfy the following property

$$P\{\hat{\delta}_k - u_k s_k \leq \delta \text{ for all } k = 1, \dots, m\} \geq 1 - \alpha.$$

---

<sup>4</sup>To see why the *lower* limit of the confidence intervals is based on the *upper* stopping boundary and vice versa, note that the confidence intervals are defined to be consistent with the group sequential decision rule. For example,  $H_0$  is rejected if  $Z_k > u_k$ , which implies that  $\hat{\delta}_k - u_k s_k > 0$ . In other words, the lower confidence limit for the treatment difference excludes 0 when the group sequential test rejects the null hypothesis of no treatment difference. Likewise,  $Z_k < l_k$  implies that the confidence interval lies completely below 0.

Repeated confidence intervals can be used in any group sequential design with both prespecified and flexible interim analyses. Unlike other types of confidence intervals that can be used only after the study has been stopped (see Section 6.2.10), repeated confidence intervals can be computed at each interim look and greatly facilitate the interpretation of interim findings.

Repeated confidence intervals are introduced to maintain the nominal coverage probability in group sequential trials. Therefore, they are wider than regular confidence intervals. The width of a repeated confidence interval depends on the selected sequential monitoring strategy. Jennison and Turnbull (1989) showed that in a clinical trial with five looks, a 90% repeated confidence interval associated with the Pocock stopping boundary is 29 percent wider than a naive confidence interval with the same coverage probability. The width ratio is constant across the five looks because the Pocock-adjusted critical values are all the same. With the O'Brien-Fleming boundary, repeated confidence intervals computed at early looks are considerably wider than regular confidence intervals. However, the width ratio decreases very quickly. In a study with five analyses, a 90% repeated confidence interval computed at the last look is only 7 percent wider than an unadjusted 90% confidence interval.

Repeated confidence intervals are easily computed in group sequential trials using PROC SEQTEST. Using the depression trial from Case study 1 as an illustration, PROC SEQTEST can be run to derive one-sided confidence intervals for the true treatment difference  $\delta$  at the individual decision points in the trial (two interim looks and final analysis).

Program 6.9 computes the lower limits of one-sided 97.5% confidence intervals for the true difference between the experimental treatment and placebo in Case study 1. To accomplish this, the `obf_teststat` data set created in Program 6.5 needs to be redefined. Specifically, we need to switch from the test statistic scale requested using `_Scale_ = 'StdZ'` to the mean difference scale (`_Scale_ = 'MLE'`). As before, the limits are derived using the O'Brien-Fleming stopping boundary generated in Program 6.1, and the boundary is passed to PROC SEQTEST using the `obf_boundary` data set. Repeated confidence intervals are requested in this program using the RCI statement.

#### **PROGRAM 6.9 Repeated confidence intervals based on the O'Brien-Fleming approach in Case study 1**

```

data obf_teststat;
  set DepTrialData;
  _Scale_= 'MLE';
  NObs= n1+n2;
  MeanDiff= mean1-mean2;
  keep _Scale_ _Stage_ NObs MeanDiff;
run;

proc seqtest boundary=obf_boundary
  data(testvar=meandiff)=obf_teststat
  nstages=3
  rci
  order=stagewise
  plots=rci;
run;

```

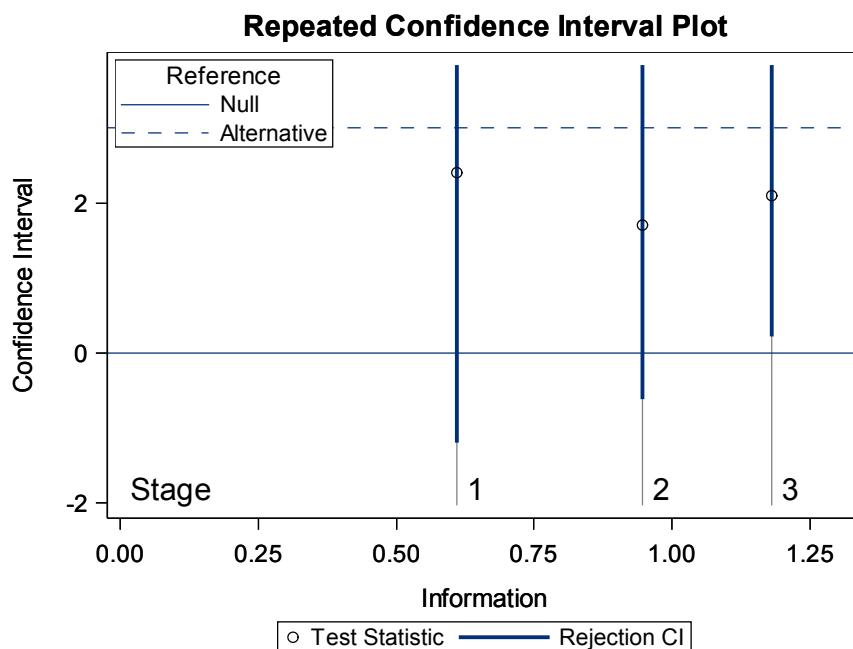
Table 6.19 lists the lower limits of one-sided repeated confidence intervals obtained at each of the three decision points in the depression trial. The lower limit is negative at both interim analyses (Stage=1 and Stage=2), indicating that the data are consistent with the null hypothesis of no treatment difference. However, the

lower limit becomes positive at the final analysis (Stage=3). Thus, with a 97.5% probability, the treatment difference is greater than 0.

**TABLE 6.19 Output from Program 6.9**

Repeated Confidence Intervals		
Stage	Parameter Estimate	Rejection Boundary Lower 97.5% CL
1	2.4	-1.2031
2	1.7	-0.6204
3	2.1	0.2235

A useful graphical summary of the repeated confidence intervals in the depression trial is presented in Figure 6.14. The figure displays the treatment differences estimated at the three decision points (open circles) as well as the lower limits of the one-sided repeated confidence intervals.

**Figure 6.14**  
Case study 1

*One-sided repeated confidence intervals for the true mean treatment difference at the three decision points.*

Table 6.19 demonstrates an important feature of repeated confidence intervals. As we can see from the table, the repeated lower limits are consistent with the O'Brien-Fleming decision rule. A repeated confidence interval includes zero if the O'Brien-Fleming test fails to detect a significant treatment difference (which was the case at the first and second interim analyses in this clinical trial) and excludes zero if the O'Brien-Fleming test rejects the null hypothesis. It is helpful to compare the adjusted confidence limits shown in Table 6.19 with the regular confidence limits used in a non-sequential setting. At the first interim look, the lower limit of a regular two-sided 95% confidence interval is 0.5120. Since this confidence interval is not properly adjusted for multiple looks, it excludes zero at this analysis. This conveys the wrong impression that the trial should have been stopped as early as the first interim look.

### 6.2.10 Estimation following sequential testing

In the previous section, we talked about repeated confidence intervals computed at each interim inspection as well as the final analysis. This section shifts the focus to inferences performed upon the trial termination and covers adjustments to final estimates of the treatment effect and associated confidence limits.

An interesting feature of maximum likelihood estimates of population parameters (e.g., sample means in the normal case or sample proportions in the binary case) is their lack of sensitivity to the underlying data collection process. Identical maximum likelihood estimates are computed from samples with a prespecified size and samples obtained in a sequentially designed experiment. For example, under normal assumptions, the maximum likelihood estimate of the population mean is the sample mean regardless of whether the sample size is fixed or random.

It has been shown that maximum likelihood estimates in a sequential setting possess the same asymptotic properties (e.g., consistency and asymptotic normality) as regular maximum likelihood estimates computed from fixed samples (see, for example, Dmitrienko and Govindarajulu, 2000). However, the non-asymptotic case is a completely different story. The distribution of maximum likelihood estimates in group sequential trials is distinctly non-normal, and the estimates themselves are not centered around the population parameters. To see why this happens, recall that an estimate of the treatment difference is computed in a sequential setting only after the trial has been stopped. At this point, the test statistic is either larger or smaller than a prespecified threshold, and, therefore, the obtained estimate is biased toward one of the two extremes. The estimate will overstate the treatment difference if the upper boundary is crossed and will have a negative bias if the lower boundary is crossed. It has been demonstrated in the literature that the bias can be quite large in magnitude and traditional estimates of the treatment effect (as well as associated confidence limits) become highly unreliable in a group sequential setting. Tsiatis, Rosner, and Mehta (1984) and Emerson and Fleming (1990) showed via simulations that the coverage probability of confidence intervals computed after a group sequential test varies substantially depending on the magnitude of the treatment difference and can be far from its nominal value.

**EXAMPLE: Group sequential trial in critically ill patients**

Van Den Berghe et al. (2001) reported the results of a clinical trial in critically ill patients that was stopped early due to a large difference in overall mortality between groups of patients treated with new and conventional therapies. The observed reduction in overall mortality was 42% with a 95% confidence interval of (22%, 62%). After a proper adjustment for sequential sampling, the estimate of the treatment effect was pulled toward 0. The adjusted reduction in overall mortality turned out to be 32% with a 95% confidence interval of (2%, 55%).

Several different approaches to computing point estimates or confidence intervals that account for the sequential nature of the testing procedure have been proposed over the last 20 years; see Siegmund (1978); Tsiatis, Rosner, and Mehta (1984); Whitehead (1986); Kim and DeMets (1987), and Emerson and Fleming (1990). In what follows, we will discuss the approach introduced by Tsiatis, Rosner, and Mehta (1984) and extended by Kim and DeMets (1987). This method possesses a number of desirable properties. For example, it produces a true confidence interval without holes, which is consistent with the associated sequential test (i.e., the interval excludes the treatment difference of 0 if the null hypothesis was rejected by the test and vice versa). The Tsiatis-Rosner-Mehta point estimates and confidence intervals are also attractive from a computational perspective. Unlike some of its competitors, the method introduced below is based on a relatively simple algorithm.

The derivation of the Tsiatis-Rosner-Mehta point estimates and confidence limits is rather complicated and will not be discussed here. In short, the derivation requires that all possible observed outcomes of a group sequential test be ordered in some meaningful manner. Tsiatis, Rosner, and Mehta (1984) considered the problem of sequential estimation of normal means and proposed to employ an ordering of the outcome space considered earlier by Siegmund (1978). Once the ordering has been specified, a bias-adjusted estimate (known as the *median unbiased estimate*) is chosen in such a way that it is less than the true treatment difference 50% of the time and greater 50% of the time. In other words, the median unbiased estimate of the true treatment difference  $\delta$  is equal to  $\hat{\delta}$  satisfying the following requirement

$$P\{\text{More extreme outcome than the observed outcome when } \delta = \hat{\delta}\} = 1/2.$$

Bias-adjusted confidence limits are derived using the well-known duality of confidence intervals and significance tests. The upper limit of a  $100(1 - \alpha)\%$  confidence interval is defined as the value  $\delta_U$  such that

$$P\{\text{More extreme outcome than the observed outcome when } \delta = \delta_U\} = 1 - \alpha/2.$$

Similarly, the lower limit is the value  $\delta_L$  such that

$$P\{\text{More extreme outcome than the observed outcome when } \delta = \delta_L\} = \alpha/2.$$

The introduced bias-adjusted confidence intervals are inherently different from repeated confidence intervals discussed in Section 6.2.9. As was stated in this section, repeated confidence intervals can be computed at any data look, including the last one. However, in general, repeated confidence intervals are best suited for assessing the size of the treatment effect at intermediate looks and are often inferior to bias-adjusted intervals constructed upon the trial termination. Recall that a repeated confidence interval is centered around the unadjusted estimate of the treatment effect. Thus, it implicitly assumes that the unadjusted estimate is reasonably close to the true population parameter, and its sample distribution is symmetric. Neither of the two assumptions holds when a group sequential trial is terminated early. Thus, a repeated confidence interval is less reliable than a bias-adjusted interval computed at the last look.

The difference between repeated and bias-adjusted confidence intervals becomes most pronounced when a decision to stop is reached very early in the trial. By the definition, repeated confidence intervals are made wide enough to ensure that intervals at *all* interim looks will contain the true treatment difference with a specified probability. If a clinical trial is terminated at the first interim analysis, the associated repeated confidence interval often turns out to be substantially wider than a bias-adjusted confidence interval that is designed to be used at a single analysis. To stress this point, Jennison and Turnbull (1990) noted that

it does appear unnecessarily conservative to use the final repeated confidence interval as a confidence interval on termination since this allows for any stopping rule whatsoever.

PROC SEQTEST supports bias-adjusted inferences at the last analysis and computes the median unbiased estimate of the treatment effect and an adjusted one-sided confidence interval at the end of a group sequential trial. It is important to mention that the estimate and confidence interval are computed only in those cases when the null hypothesis is rejected, i.e., the experimental treatment is found to be effective. As in the case of repeated confidence intervals, it is theoretically possible to perform adjusted inferences after the futility stopping boundary has been crossed. However, these inferences might not be of much practical interest in a clinical trial setting.

Program 6.9 introduced in Section 6.2.9 also computes the bias-adjusted point estimate and the lower limit of a two-sided 95% confidence interval for the true treatment difference at the last analysis in Case study 1. The inferences are performed using the O'Brien-Fleming stopping boundary generated earlier in Program 6.1. The resulting median unbiased estimate as well as the lower confidence limit are shown in Table 6.20. It is instructive to compare the derived median unbiased estimate to the regular (observed) mean treatment difference computed at the final analysis. As shown in the table, the observed mean treatment difference (MLE) at the last analysis is 2.1, and it is more extreme than the bias-adjusted estimate (Median Estimate), i.e., 2.0027. As expected, removing the bias caused by sequential sampling shifts the estimate of the treatment effect toward zero. Further, the bias-adjusted lower confidence limit displayed in Table 6.20 (0.4405) is different from the lower limit of the repeated confidence interval at the final analysis given in Table 6.19 (0.2235). As was explained above, the bias-adjusted limit is likely more reliable in this situation than its counterpart since it better accounts for the possibly non-normal distribution of the treatment difference at the last look.

**TABLE 6.20 Output from Program 6.9**

Parameter	Stopping Stage	Parameter Estimates Stagewise Ordering			Median Estimate	Lower 95% CL
		MLE	p-value for H0:Parm=0			
meandiff	3	2.1	0.0178	2.0027	0.4405	

### 6.2.11 Summary

Sequential designs have long become part of the standard arsenal of statistical tools used in clinical trials. This section reviewed a popular approach to the design and monitoring of group sequential trials known as the repeated significance testing approach.

First, we provided a review of popular group sequential plans (O'Brien-Fleming and Pocock plans) and demonstrated how to design group sequential trials for testing efficacy or simultaneous testing of efficacy and futility of an experimental treatment. PROC SEQDESIGN can be used to set up these and other group sequential designs.

The section also discussed interim monitoring based on the flexible error spending approach. The main advantage of this approach is its ability to allow for unplanned interim looks. Therefore, it should not be surprising that it has become the main tool for the sequential analysis of clinical trial data. It is important to keep the following two points in mind when applying the error spending approach in a clinical trial setting:

- The maximum sample size must be specified prior to the first interim analysis.
- The monitoring plan cannot be affected by the outcomes observed in the trial. Although it has been shown that the error spending approach is quite robust to data-dependent changes in the frequency of interim analyses, clinical researchers need to avoid such practices. Requesting additional looks at the data after a nearly significant  $p$ -value was observed or other data-driven decisions will inflate the overall Type I error rate.

The error spending approach is implemented in PROC SEQTEST and supports data monitoring for detecting early evidence of efficacy. When performing efficacy monitoring, it is critical to properly adjust the point estimates and confidence

intervals computed in group sequential trials. The adjustments reflect the fact that the trial was conducted in a sequential manner. The section reviewed two types of adjustment methods:

- **Repeated confidence intervals** provide a useful summary of the variability of the estimated treatment difference at each interim look. Repeated confidence intervals are defined to ensure that all confidence intervals will contain the true treatment difference with a specified probability. If a clinical trial is stopped early, a repeated confidence interval computed at the last analysis will be wider (and often substantially wider) than a confidence interval specifically designed to be used at the end of a group sequential trial.
- **Point estimates and confidence intervals at termination.** Naive estimates of the treatment effect and associated confidence intervals used in traditional designs are misleading in trials that employ sequential designs because these estimates tend to be biased toward more extreme values of the treatment difference. In order to alleviate this problem, point estimates and confidence limits following termination are constructed by removing the bias caused by sequential sampling. For instance, if an estimate of the treatment difference is biased upward, it will be shrunk toward zero using a suitable bias-adjustment procedure.

Repeated confidence intervals as well as adjusted point estimates and confidence limits at termination are easy to compute using PROC SEQTEST.

## 6.3 Stochastic curtailment tests

---

We have pointed out repeatedly throughout the chapter that pivotal clinical trials are routinely monitored to detect early evidence of beneficial or harmful effects. Section 6.2 introduced an approach to the design and analysis of group sequential trials based on repeated significance testing. This section looks at sequential testing procedures used in a slightly different context. Specifically, it reviews recent advances in the area of *stochastic curtailment* with emphasis on procedures for futility monitoring.

To help understand the difference between repeated significance and stochastic curtailment tests, recall that repeated significance testing is centered around the notion of error spending. The chosen error spending strategy determines the characteristics of a sequential test (e.g., the shape of stopping boundaries) and ultimately drives the decision-making process. In the stochastic curtailment framework, decision making is tied directly to the final trial outcome. A decision to continue the trial or curtail sampling at each interim look is based on the likelihood of observing a positive or negative treatment effect if the trial were to continue to the planned end. Another important distinction is that stochastic curtailment methods are aimed toward predictive inferences, whereas repeated significance tests focus on currently available data.

Stochastic curtailment procedures (especially futility stopping rules) are often employed in clinical trials with a traditional design without an explicit adjustment for repeated statistical analyses. It is not unusual to see a stochastic curtailment test used in a *post hoc* manner, for example, a futility rule might be adopted after the first patient visit.

This section reviews three popular types of stochastic curtailment tests that employ different definitions of predicted probability of success to construct futility stopping rules in clinical trials. Section 6.3.1 introduces frequentist methods commonly referred to as conditional power methods. We will discuss the conditional power test originally defined by Lan, Simon, and Halperin (1982) and Halperin et al.

(1982) as well as its extensions proposed by Pepe and Anderson (1992) and Betensky (1997). This section also discusses mixed Bayesian-frequentist methods (Herson, 1979; Spiegelhalter, Freedman, and Blackburn, 1986). Section 6.3.2 introduces fully Bayesian methods considered, among others, by Geisser (1992), Geisser and Johnson (1994), Johns and Anderson (1999), and Dmitrienko and Wang (2006). See also Lee and Liu (2008). Finally, Section 6.3.3 briefly describes other applications of stochastic curtailment tests in clinical trials with data-driven decision rules, including rules to modify the trial's sample size or select a subpopulation of patients based on the results observed at an interim analysis.

## Cases studies

---

The following two cases studies will be used to illustrate popular stochastic curtailment tests reviewed in this section. In both clinical trial examples we are interested in setting up a futility rule to help assess the likelihood of a positive trial outcome given the results observed at each interim look.

**EXAMPLE: Case study 3 (Severe sepsis trial with a binary endpoint)**

Consider a Phase II clinical trial in patients with severe sepsis conducted to compare the effect of an experimental treatment on 28-day all-cause mortality to that of placebo. The trial design included a futility monitoring component based on monthly analyses of the treatment effect on survival. The total sample size of 424 patients was computed to achieve 80% power of detecting a 9% absolute improvement in the 28-day survival rate at a one-sided significance level of 0.1. This calculation was based on the assumption that the placebo survival rate is 70%.

The trial was discontinued after approximately 75% of the projected number of patients had completed the 28-day study period because the experimental treatment was deemed unlikely to demonstrate superiority to placebo at the scheduled termination point. Table 6.21 provides a summary of the mortality data at six interim analyses conducted prior to the trial's termination.

**TABLE 6.21** Survival data in the severe sepsis trial

Interim analysis	Treatment arm			Placebo arm		
	n	Number of survivors	Survival rate	n	Number of survivors	Survival rate
1	55	33	60.0%	45	29	64.4%
2	79	51	64.6%	74	49	66.2%
3	101	65	64.4%	95	62	65.3%
4	117	75	64.1%	115	77	67.0%
5	136	88	64.7%	134	88	65.7%
6	155	99	63.9%	151	99	65.6%

**EXAMPLE: Case study 4 (Generalized anxiety disorder trial with a continuous endpoint)**

A small proof-of-concept trial in patients with generalized anxiety disorder will serve as an example of a trial with a continuous outcome variable. In this trial, patients were randomly assigned to receive an experimental treatment or placebo for 6 weeks. The efficacy profile of the treatment was evaluated using the mean change from baseline in the Hamilton Anxiety Rating Scale (HAMA) total score. As in Case study 1, a larger reduction in the HAMA total score indicates a beneficial effect. The total sample size was 82 patients (41 patients were to be enrolled in each of the two trial arms). This sample size provides approximately 80% power

to detect a treatment difference of 4 in the mean HAMA total score change. The power calculation is based on a one-sided test with an  $\alpha$ -level of 0.1 and assumes a standard deviation of 8.5.

The trial was not intended to be stopped early in order to declare efficacy. However, HAMA total score data were monitored in this trial to detect early signs of futility or detrimental effects. Table 6.22 summarizes the results obtained at three interim looks (positive quantities are presented in this table to show improvement, i.e., reduction, in the HAMA total score from baseline).

**TABLE 6.22** **HAMA total score data in the generalized anxiety disorder trial**

Interim analysis	n	Treatment arm	Placebo arm	
		Mean HAMA improvement (SD)	n	Mean HAMA improvement (SD)
1	10	9.2 (7.3)	11	8.4 (6.4)
2	20	9.4 (6.7)	20	8.1 (6.9)
3	31	8.9 (7.4)	30	9.1 (7.7)

### 6.3.1 Futility rules based on conditional and predictive power

This section discusses applications of stochastic curtailment tests to futility monitoring in clinical trials. As noted in Section 6.2, futility monitoring plays a very important role in clinical trials. In a nut shell, a futility test causes the clinical trial to be stopped as soon as it becomes clear that a negative outcome is inevitable and that it is no longer worthwhile continuing the trial to its completion. As pointed out by Ware, Muller, and Braunwald (1985),

... early termination for futility could reduce the enormous expenditures of resources, human and financial, involved in the conduct of trials that ultimately provide a negative answer regarding the value of a medical innovation.

For this reason, methods for assessing the strength of evidence against the alternative hypothesis, known as *futility stopping rules*, are used throughout the drug development cycle. They help quickly eliminate weak candidates in early proof-of-concepts studies and minimize the number of patients exposed to ineffective treatments for treating conditions that might be fatal or have irreversible outcomes.

Futility stopping rules are commonly used in Phase II and Phase III trials and, unlike efficacy stopping rules discussed in Section 6.2, futility assessments are often performed multiple times throughout the trial. It is important to bear in mind that interim analyses conducted for the sole purpose of futility monitoring have no impact on Type I error rate control (but the Type I error rate will be deflated with frequent futility assessments). In addition, as explained in Section 6.2, futility stopping rules are typically set up in a non-binding way in confirmatory clinical trials, which means that they might be overridden by the trial's sponsor. For this reason, futility-based decision rules are often applied in a flexible way, e.g., the schedule of futility assessments might not be predefined, and parameters of these decision rules might be updated based on the observed data. However, to prevent power loss, it is helpful to evaluate the impact of futility tests on the trial's operating characteristics, and, in fact, trial sponsors are encouraged to select the timing of futility assessments and parameter of futility stopping rules in an efficient way. Applications of clinical trial optimization approaches aimed at identifying optimal futility stopping rules were discussed in Dmitrienko et al. (2016). See also Dmitrienko and Pulkstenis (2017).

Other applications of stochastic curtailment tests in clinical trials, including their use in adaptive clinical trials with sample size adjustment and complex selection rules, are presented in Section 6.3.3.

## Conditional power

Conditional power provides the most basic approach to defining predicted probability of success in clinical trials with multiple decision points. To introduce the concept of conditional power, consider a two-arm clinical trial (experimental treatment versus control) with a normally distributed primary endpoint. A balanced design with the same number of patients enrolled in the control and treatment arms will be assumed. The planned sample size per trial arm at the final decision point is equal to  $N$  patients. Assume that an interim look at the data will be taken after  $n$  patients per arm complete the trial. Let  $t$  denote the information fraction at the interim analysis, i.e.,  $t = n/N$ . Finally,  $\delta$  and  $\sigma$  denote the true mean treatment difference and common standard deviation of the primary endpoint, respectively, and  $\theta = \delta/\sigma$  is the true effect size (standardized treatment difference).

Let  $Z_t$  denote the test statistic computed at the interim look, i.e.,

$$Z_t = \frac{\hat{\delta}_t}{\hat{\sigma}_t \sqrt{2/n}},$$

where  $\hat{\delta}_t$  is the estimated mean treatment difference and  $\hat{\sigma}_t$  is the pooled sample standard deviation at the interim analysis. Similarly,  $Z_1$  denotes the test statistic at the planned end of the trial, which corresponds to the information fraction  $t = 1$ , i.e.,

$$Z_1 = \frac{\hat{\delta}_1}{\hat{\sigma}_1 \sqrt{2/N}}.$$

The conditional power test proposed by Lan, Simon, and Halperin (1982) is based on an intuitively appealing idea of predicting the distribution of the final outcome given the data already observed in the trial. If the interim prediction indicates that the trial outcome is unlikely to be positive, ethical and financial considerations will suggest an early termination the trial. Mathematically, the decision rule can be expressed in terms of the probability of a statistically significant outcome, i.e.,  $Z_1 > z_{1-\alpha}$ , conditional on the interim test statistic  $Z_t$ . This conditional probability is termed *conditional power*. The trial is stopped as soon as the conditional power falls below a pre-specified value  $1 - \gamma$ , where  $\gamma$  is known as the *futility index* (Ware, Muller and Braunwald, 1985). Smaller values of  $\gamma$  are associated with a higher likelihood of early stopping and, most commonly,  $\gamma$  is set to a value between 0.8 and 1.

It was shown by Lan, Simon, and Halperin (1982) that the outlined futility stopping rule relies on a simple closed-form expression. Suppose first that the true effect size  $\theta$  is known. It is easy to verify that the test statistics  $Z_t$  and  $Z_1$  follow a bivariate normal distribution, and, as a result, the conditional distribution of  $Z_1$  given  $Z_t$  is normal with the following parameters:

$$\text{mean} = \sqrt{t}Z_t + \theta(1-t)\sqrt{\frac{N}{2}} \text{ and variance} = 1 - t.$$

Therefore, the conditional power function at the interim analysis, denoted by  $P_t(\theta)$ , is given by

$$P_t(\theta) = P(Z_1 > z_{1-\alpha}|Z_t) = \Phi \left( \sqrt{\frac{t}{1-t}}Z_t + \theta\sqrt{\frac{(1-t)N}{2}} - \frac{z_{1-\alpha}}{\sqrt{1-t}} \right),$$

where, as before,  $z_{1-\alpha}$  denotes the  $100(1 - \alpha)$ th percentile of the standard normal distribution and  $\Phi(x)$  is the cumulative probability function of the standard normal distribution.

A stopping rule can now be formulated in terms of the computed conditional power, i.e.,

Terminate the trial due to lack of efficacy the first time  $P_t(\theta) < 1 - \gamma$ ,

or in terms of the test statistic computed at the interim analysis  $Z_t$ :

Terminate the trial due to lack of efficacy the first time  $Z_t < l_t$ ,

where the adjusted critical value  $l_t$  is given by

$$l_t = \sqrt{t} z_{1-\alpha} + \sqrt{\frac{1-t}{t}} z_{1-\gamma} - \theta(1-t) \sqrt{\frac{N}{2t}}.$$

This approach presented above is easy to extend to clinical trials with a binary primary endpoint. Specifically, let  $p_1$  and  $p_2$  denote the true event rates in the treatment and control arms, and assume that a higher rate indicates a beneficial effect. In this case, conditional power  $P_t(\theta)$  is computed as shown above with the effect size given by

$$\theta = \frac{p_1 - p_2}{\sqrt{\bar{p}(1-\bar{p})}},$$

where  $\bar{p}$  is the average response rate in the two trial arms, i.e.,  $\bar{p} = (p_1 + p_2)/2$ .

A slightly different formula needs to be used to compute conditional power in trials with a time-to-event primary endpoint, e.g., a survival endpoint (see Jennison and Turnbull, 1990, for more details). In this setting, conditional power is given by

$$P_t(\theta) = \Phi \left( \sqrt{\frac{t}{1-t}} Z_t + \theta \sqrt{\frac{(1-t)N}{4}} - \frac{z_{1-\alpha}}{\sqrt{1-t}} \right),$$

where  $\theta$  is the logarithm of the true hazard ratio.

In addition, this approach is easy to extend to clinical trials with an unbalanced design. Let  $r$  denote the randomization ratio, e.g.,  $r = 2$  if patients are randomly allocated to the control and treatment arms using a 1:2 ratio. In a trial with a normally distributed endpoint, conditional power is given by

$$P_t(\theta) = P(Z_1 > z_{1-\alpha} | Z_t) = \Phi \left( \sqrt{\frac{t}{1-t}} Z_t + \theta \sqrt{\frac{(1-t)rN}{1+r}} - \frac{z_{1-\alpha}}{\sqrt{1-t}} \right),$$

## Choice of effect size or treatment difference

The conditional power function  $P_t(\theta)$  introduced above relies on the assumption that the true effect size  $\theta$  is known to the trial's sponsor. Since this is clearly not the case in real-world clinical trials, the sponsor needs to select an appropriate value of  $\theta$  when computing conditional power. This value will define the assumed distribution of the future data that will be generated after the interim analysis. Lan, Simon and Halperin (1982) and Halperin et al. (1982) considered futility stopping rules with the conditional power function evaluated at the value of  $\theta$  corresponding to the alternative hypothesis of beneficial treatment effect. It has been noted in the literature that this approach to defining conditional power might lead to spurious results, and, as a consequence, the corresponding conditional power test does not

always serve as a reliable stopping criterion. For example, Pepe and Anderson (1992) reviewed the performance of the Lan-Simon-Halperin test under a scenario when interim data do not support the alternative hypothesis. They showed that this test tends to produce an overly optimistic probability of a positive trial outcome when the observed treatment effect is substantially different from the hypothesized one.

Since failure to account for emerging trends might undermine the utility of the simple conditional power method, Lan and Wittes (1988) proposed to evaluate the conditional power function at a data-driven value. Pepe and Anderson (1992) studied a conditional power test using the value of the treatment difference that is slightly more optimistic than the maximum likelihood estimate of the treatment difference computed at the interim analysis. Further, using the Pepe-Anderson test as a starting point, Betensky (1997) introduced a family of adaptive conditional power tests in which the assumed value of the mean treatment difference (e.g., the mean difference under the alternative hypothesis) is replaced with an appropriate sample estimate obtained at the interim analysis. The approaches to defining conditional power achieve greater efficiency than the simple conditional power method that relies on the alternative hypothesis by “letting the data speak for themselves.”

As a quick illustration, consider a clinical trial with a normally distributed primary endpoint. Assuming that the standard deviation  $\sigma$  is known, let  $P_t(\delta)$  denote conditional power as a function of the true mean treatment difference  $\delta$ . Betensky (1997) proposed to replace  $\delta$  in  $P_t(\delta)$  with

$$\hat{\delta}_t + cSE(\hat{\delta}_t).$$

Here  $\hat{\delta}_t$  is the maximum likelihood estimate of  $\delta$  from the data available at the interim look corresponding to the information fraction  $t$ .  $SE(\hat{\delta}_t)$  denotes its standard error, and  $c$  is a positive parameter that determines the shape of the stopping boundary. The  $c$  parameter equals 1 in the Pepe-Anderson test and Betensky (1997) recommended to let  $c = 2.326$  (i.e., the 99th percentile of the standard normal distribution) to better control the amount of uncertainty about the final trial outcome.

It will be shown below how PROC SEQTEST can be used to perform conditional power calculations under the original proposal due to Lan, Simon, and Halperin (1982) as well as alternative approaches developed in Betensky (1997).

## **Relationship between repeated significance and conditional power tests**

It is worth noting that tests based on conditional power are closely related to repeated significance tests used in group sequential trials that were discussed in Section 6.2. Jennison and Turnbull (1990) pointed out that a special case of the Lan-Simon-Halperin test introduced above is a continuous-time equivalent of the O’Brien-Fleming group sequential test. Davis and Hardy (1990) showed that the stopping boundary of the Lan-Simon-Halperin test for efficacy monitoring with  $\gamma = 0.5$  is virtually equivalent to the stopping boundary of the O’Brien-Fleming test. Likewise, if we construct the O’Brien-Fleming boundary for testing lack of treatment benefit, it will be very close to the stopping boundary of the Lan-Simon-Halperin test with  $\gamma = 0.5$ . Note, however, that the futility index  $\gamma$  is typically set at the 0.8 or higher level in clinical trials. This implies that a test for futility, based on the O’Brien-Fleming method (and other repeated significance testing methods), is more likely to trigger an early stopping due to lack of efficacy compared to the Lan-Simon-Halperin test.

## Predictive power

An important feature of methods based on conditional power is that they rely on a purely frequentist framework in the sense that the estimated effect size used for generating future data is treated as a fixed value. The observed treatment difference and its variability support predictions after the interim analysis. However, no adjustment is made to account for the associated prediction error. By contrast, *predictive power* tests discussed below adopt more of a Bayesian approach to computing the predicted probability of success. The predictive power methodology involves averaging the conditional power function (frequentist concept) with respect to the posterior distribution of the treatment effect given the observed data (Bayesian concept). For this reason, this methodology is often referred to as the *mixed Bayesian-frequentist* methodology. By using the posterior distribution we can improve our ability to quantify the uncertainty about future data, for instance, to better account for the possibility that the current trend might be reversed.

To define the predictive power approach, consider again a two-arm clinical trial with  $N$  patients per trial arm. An interim analysis is conducted after  $n$  patients completed the trial in each arm. Let  $Z_t$  and  $Z_1$  denote the test statistics computed at the interim and final analyses, respectively. As before,  $t$  denotes the information fraction at this interim look, i.e.,  $t = n/N$ . Assume for the sake of illustration that the primary endpoint in the trial is normally distributed. Let  $\delta$  denote the unknown true mean treatment difference and  $\sigma$  denote the known common standard deviation in the two trial arms. The futility rule is expressed in terms of the following quantity

$$P_t = \int P_t(\delta) f(\delta|Z_t) d\delta,$$

where  $P_t(\delta)$  is conditional power as a function of the true treatment difference  $\delta$  and  $f(\delta|Z_t)$  is the posterior density of  $\delta$  given the observed data at the interim look. The resulting quantity,  $P_t$ , can be thought of as the average conditional power and is termed *predictive power*. The trial is terminated at the interim analysis due to lack of treatment benefit if predictive power is less than  $1 - \gamma$  for some prespecified futility index  $\gamma$  and continues otherwise. Using similar ideas, predictive power can also be derived in clinical trials with primary endpoints that do not follow a normal distribution.

As a side note, it is worth mentioning that we can also set up a predictive power test with a weight function based on the prior distribution of the true treatment difference  $\delta$ . This approach is similar to the basic conditional power approach in that it puts too much weight on original assumptions and ignores the interim data. Thus, it should not be surprising that predictive power tests based on the prior distribution are inferior to their counterparts relying on the posterior distribution of the treatment effect (Bernardo and Ibrahim, 2000).

Closed-form expressions for computing predictive power can be found in Dmitrienko and Wang (2006). PROC SEQTEST supports predictive power calculations under the assumption that the prior distribution of the true treatment difference is non-informative. Predictive power calculations with informative priors can be performed using the %BayesFutilityCont and %BayesFutilityBin macros introduced in Section 6.3.2.

## Futility stopping rules based on conditional and predictive power in Case study 3

The Phase II trial for the treatment of severe sepsis will be used to illustrate the process of computing conditional and predictive power and application of futility

stopping rules in clinical trials. Program 6.10 computes conditional power at the very first look in the severe sepsis trial (as shown in Table 6.21, this look corresponds to the information fraction of 0.24). As the first step, PROC SEQDESIGN is used to set up an overall framework for computing conditional power. As shown in the program, the primary endpoint in this trial is binary, and the treatment effect will be evaluated using a two-sample test for proportions. The event rate (28-day survival rate) under the null hypothesis of no effect is defined using `nullprop=0.70`. The treatment difference (i.e., the difference in the survival rates in the two trial arms) under the alternative hypothesis is defined using `altref=0.09`. In addition, the number of decision points is specified (`nstages=2`) along with the information fractions at these decision points (`info=cum(0.24 1)`). Note that the `ceiladjdesign=include` option in PROC SEQDESIGN forces the group sequential design to have an integer-valued rather than a fraction sample size at the interim look.

#### **PROGRAM 6.10 Conditional power tests at the first interim look in Case study 3**

```

proc seqdesign altref=0.09;
  design method=errfuncobf
  nstages=2
  info=cum(0.24 1)
  alt=upper
  stop=reject
  alpha=0.1
  beta=0.2;
  samplesize model(ceiladjdesign=include)
    =twosamplefreq(nullprop=0.70 test=prop ref=avgprop);
  ods output adjustedboundary=obf_boundary;
run;

data sevsep1;
  input _stage_ treatment nobs nsurv;
  datalines;
  1 1 55 33
  1 0 45 29
run;

proc genmod data=sevsep1;
  model nsurv/nobs= treatment / dist=bin link=identity;
  ods output nobs=ntotal parameterestimates=parmest;
run;

data ntotal;
  set ntotal;
  nobs= n;
  if (label='Number of Trials');
  keep nobs;
run;

data parms;
  set parmest;
  if parameter='treatment';
  _scale_='MLE';
  keep _scale_ parameter estimate;
run;

data parms;
  _stage_=1;
  merge ntotal parms;

```

```

run;

proc seqtest boundary=obf_boundary
  parms(testvar=treatment infovar=nobs)=parms
  boundaryadj=errfuncobf
  boundarykey=alpha
  nstages=2
  order=stagewise
  condpower(type=finalstage);
run;

```

Further, the number of patients and their survival outcomes in the two arms from Table 6.21 are defined in the `sevsep1` data set and the two-sample test for proportions is carried out using PROC GENMOD. The treatment difference at the first interim analysis is  $\hat{\delta} = -0.044$  with the standard error given by  $SE(\hat{\delta}) = 0.0972$ . (Note that the treatment difference is negative since the experimental treatment results in an increased mortality rate.) The resulting estimates are saved in the `parms` data set and are passed to PROC SEQTEST to compute conditional power. It is worth noting that conditional power is defined as the conditional probability of a statistically significant treatment effect at the final analysis. Thus, the parameter of the conditional power test is set to `type=finalstage`.

Table 6.23 lists multiple values of conditional power computed by PROC SEQTEST. The first row in this table (Ref=MLE) corresponds to conditional power evaluated at the observed treatment difference (i.e.,  $\delta = \hat{\delta} = -0.0444$ ) and indicates that the predicted probability of success is extremely low (0.38%). Note that the CRef value is computed relative to the treatment difference under the alternative hypothesis, i.e.,  $-0.0444/0.09 = -0.4938$ . The second row (Ref=NULL) shows conditional power, which is obtained under the assumption that the future data will be generated from the null hypothesis of no treatment effect, i.e.,  $\delta = 0$ . The resulting value is also quite low (4%). The probabilities in the top two rows of Table 6.23 suggest that the trial should be terminated due to lack of efficacy. Conditional power in the fourth row (Ref=Alternative) is derived using the Lan-Simon-Halperin approach, i.e., with the treatment difference set to its value under the alternative hypothesis (i.e.,  $\delta = 0.09$ ). In this case, the future data are consistent with the hypothesis of a strong treatment effect. The resulting conditional power indicates that, despite a negative treatment difference at the first interim look, there is still a 55% chance that the trend will be reversed by the completion of the trial. Lastly, the values of conditional power shown in the third and fifth rows can be used to inform the decision-making process. For example, conditional power in the third row (CRef=0.5) is evaluated at the mid-point between the null and alternative hypotheses--in other words, under the assumption that the 28-day survival rate in the treatment arm is 74.5%.

**TABLE 6.23 Output from Program 6.10**

Stopping Stage	Conditional Power Information Reference = CRef * (Alt Reference)			
	MLE	Reference	CRef	Conditional Power
1	-0.0444	MLE	-0.4938	0.0038
1	-0.0444	Null	0	0.0400
1	-0.0444		0.50	0.2057
1	-0.0444	Alternative	1.00	0.5428
1	-0.0444		1.50	0.8501

To summarize the results presented in Table 6.23, conditional power evaluated under the observed treatment difference (Ref=MLE) or under the null hypothesis

of no effect (Ref=Null) is much lower than conditional power evaluated under the alternative (Ref=Alternative). This observation has a simple explanation. Since the true treatment difference for the future observations is assumed to be smaller in magnitude, there is a lower chance that the negative treatment difference will get reversed by the end of the trial and thus the conditional power of a positive trial outcome will also be lower. We can see from Table 6.23 that making an overly optimistic assumption about the size of the treatment effect delays the decision to terminate the trial and can potentially cause more patients to be exposed to an ineffective therapy. For this reason, in the absence of strong *a priori* evidence, it might be inappropriate to use the alternative hypothesis for generating future data in conditional power calculations.

Program 6.10 focused on several most commonly used definitions of conditional power, and it is shown in Program 6.11 how PROC SEQTEST can be used to support the general data-driven approach to defining conditional power developed by Betensky (1997). As explained above, Betensky (1997) proposed to plug the following estimate into the expression into the conditional power function:

$$\hat{\delta} + cSE(\hat{\delta}),$$

where  $c$  is a predefined constant. As this constant increases, increasingly more optimistic assumptions are made about the true treatment effect in the data to be observed after the interim analysis. Pepe and Anderson (1992) proposed to set  $c = 1$  and, to illustrate this approach to defining conditional power, recall that the observed treatment difference and its standard error at the first interim look are given by

$$\hat{\delta} = -0.0444, SE(\hat{\delta}) = 0.0972.$$

To compute conditional power in PROC SEQTEST with  $\delta = \hat{\delta} + SE(\hat{\delta}) = 0.0528$ , a reference value needs to be computed relative to the treatment difference assumed under the alternative hypothesis, i.e., relative to 0.09. This means that the reference value needs to be set to 0.5867, i.e., `cref=0.5867`. The PROC SEQTEST code that performs this calculation is shown in Program 6.11.

#### **PROGRAM 6.11 Conditional power test based on the Pepe-Anderson approach at the first interim look in Case study 3**

```
proc seqtest boundary=obf_boundary
  parms(testvar=treatment infovar=nobs)=parms
  boundaryadj=errfuncobf
  boundarykey=alpha
  nstages=2
  order=stagewise
  condpower(cref=0.5867 type=finalstage);
run;
```

Table 6.24 shows the value of conditional power based on the Pepe-Anderson approach at the first interim look in the severe sepsis trial. Even though the underlying assumption used in this calculation is more optimistic than the assumption that

**TABLE 6.24 Output from Program 6.11**

Conditional Power Information				
Reference = CRef * (Alt Reference)				
Stopping Stage	MLE	Reference Ref	CRef	Conditional Power
1	-0.0444	MLE	-0.4938	0.0038
1	-0.0444	0.5867	0.2545	

the data after the interim will be fully consistent with the data observed prior to the interim (i.e. Ref=MLE) the resulting predicted probability of success at the end of the trial is still very low (25.5%). This again suggests that it would be advisable to terminate this Phase II trial due to futility at the very first interim analysis.

Program 6.12 builds upon Program 6.10 to perform conditional power and predictive power calculations at all six predefined interim analyses in Case study 3. These calculations will help us understand how the different approaches to predicting the probability of success use information on treatment effect trends over multiple looks. Conditional power is computed in Program 6.12 using the following three methods:

- Conditional power is evaluated under the alternative hypothesis of a beneficial effect.
- Conditional power is evaluated using the observed value of the treatment difference at the interim analysis.
- Conditional power is evaluated under the null hypothesis of no treatment effect.

### **PROGRAM 6.12 Conditional and predictive power tests at six interim looks in Case study 3**

```

data sevsep;
  input _stage_ treatment nobs nsurv;
  datalines;
    1 1 55 33
    1 0 45 29
    2 1 79 51
    2 0 74 49
    3 1 101 65
    3 0 95 62
    4 1 117 75
    4 0 115 77
    5 1 136 88
    5 0 134 88
    6 1 155 99
    6 0 151 99
  run;

%SeqPower(data=sevsep, analysis=1, fraction=0.24, out=power1);
%SeqPower(data=sevsep, analysis=2, fraction=0.36, out=power2);
%SeqPower(data=sevsep, analysis=3, fraction=0.46, out=power3);
%SeqPower(data=sevsep, analysis=4, fraction=0.55, out=power4);
%SeqPower(data=sevsep, analysis=5, fraction=0.64, out=power5);
%SeqPower(data=sevsep, analysis=6, fraction=0.72, out=power6);

data power;
  set power1-power6;
run;

proc print data=power noobs;
  var analysis mle cmle cnnull calt ppower;
run;

```

In addition, predictive power is evaluated at each of the six interim analyses in the severe sepsis trial.

To support an efficient evaluation of conditional and predictive power, Program 6.12 relies on the %SeqPower macro which performs conditional and predictive

power calculations at a single interim analysis using the same steps that were used in Program 6.10. This macro has the following arguments:

- **Data** is the name of the input data set.
- **Analysis** is the number of the current interim analysis.
- **Fraction** is the information fraction at the current interim analysis.
- **Out** is the name of the data set where the conditional and predictive power values will be stored.

The numbers of patients included in the interim analysis data set at each look as well as the number of survivors in each trial arm are included in the **sevsep** data set. This data set is passed to **%SeqPower** to carry out the conditional power and predictive power tests at each interim analysis in the trial.

Table 6.25 provides a summary of conditional power and predictive power calculations at the six interim looks performed by Program 6.12. To facilitate a review of the results, suppose that we are interested in applying a futility stopping rule based on a 20% threshold for the predicted probability of success. In other words, the trial will be terminated due to lack of efficacy as soon as the predicted probability of success falls below 20%.

**TABLE 6.25** Output from Program 6.12

Interim look	MLE	Conditional power			Predictive power
		Ref=MLE	Ref=Null	Ref=Alternative	
1	-0.0444	0.0038	0.0400	0.5428	0.0975
2	-0.0166	0.0176	0.0367	0.4662	0.1035
3	-0.0091	0.0195	0.0283	0.3720	0.0816
4	-0.0285	0.0016	0.0067	0.1568	0.0158
5	-0.0097	0.0053	0.0079	0.1398	0.0222
6	-0.0169	0.0006	0.0013	0.034	0.0035

When conditional power was evaluated under the alternative hypothesis (Ref=Alternative), a very optimistic assumption was made regarding the treatment effect in the future data. Despite the fact that the treatment effect was negative at each look, this assumption virtually excluded the possibility of an early stopping due to futility. The conditional power values were, in fact, very high at the first three interim look. Although the conditional power decreases steadily, it remained above the prespecified 20% threshold until the fourth interim look. Based on this criterion, the futility stopping rule would be met only after the mid-point in this trial (recall that the information fraction at the fourth interim analysis is 0.55).

The conditional power values based on the observed treatment difference (Ref=MLE) and null hypothesis (Ref=Null) were much lower compared to the first definition of conditional power. The conditional power values in both cases were below the 20% threshold at the very first look, which suggested that the data at that look did not support the alternative hypothesis. The futility stopping rule would be met at the first interim analysis. This early rejection of the alternative hypothesis is consistent with the observation that the experimental treatment is unlikely to improve survival and might even have a detrimental effect.

The conclusion based on predictive power is similar to that based on the data-driven definition of conditional power. Using predictive power, the futility stopping rule would again be met at the first interim analysis. Also, considering the predictive power values listed in the table, it is important to note that predictive power was computed based on a non-informative prior. In general, a non-informative or a weak prior increases the chances of an early stopping because the prior is easily dominated by the observed data, and the resulting predictive power begins to approach the

value of conditional power based on the treatment effect estimated from the interim data. This trend is illustrated in Table 6.25. The predictive power values were indeed fairly close to the conditional power values derived from the observed treatment difference (Ref=MLE). By contrast, if a strong prior centered around the alternative hypothesis was assumed, the future data would be essentially generated from this hypothesis. As a result, the predictive power method with a strong optimistic prior would produce overly optimistic predictions about the final trial outcome. Clinical researchers need to keep this relationship in mind when selecting futility stopping rules based on predictive power.

### Type II error rate control

The last remark in this section is concerned with the effect of futility monitoring on the Type II error rate or power in a clinical trial. It was stressed before that, due to the possibility of early stopping in favor of the null hypothesis, any futility monitoring strategy results in power loss, and it is important to assess the amount of Type II error rate inflation. Toward this end, Lan, Simon, and Halperin (1982) derived an exact lower bound for the power of their test. They showed that the power under the alternative hypothesis cannot be less than  $1 - \beta/\gamma$ . For example, the power of the severe sepsis trial with  $\beta = 0.2$  employing the Lan-Simon-Halperin futility rule will always exceed 77.8% if  $\gamma = 0.9$  and 71.4% if  $\gamma = 0.7$ , regardless of the number of interim analyses.

Although it is uncommon to do so, we can theoretically use the obtained lower bound to re-calculate the sample size and bring the power back to the desired level. After the sample size has been adjusted, we can apply the Lan-Simon-Halperin futility rule an arbitrary number of times without compromising the trial's operating characteristics. For instance, in order to preserve the Type II error rate at the 0.2 level, the power of the severe sepsis trial needs to be set to  $1 - \beta\gamma$ , which is equal to 82% if  $\gamma = 0.9$  and 86% if  $\gamma = 0.7$ . To see how this affects the trial's size, recall that the original traditional design with a one-sided  $\alpha$ -level of 0.1 and 80% power requires 424 patients. To achieve 82% power, the number of patients will have to be increased to 454 (7% increase). Similarly, 524 patients (24% increase) will need to be enrolled in the trial to protect the Type II error probability if  $\gamma = 0.7$ .

We should remember that the exact lower bound for the overall power is rather conservative and is achieved only in clinical trials with a very large number of interim looks (especially if  $\gamma$  is less than 0.7). If re-computing the sample size appears feasible, we can also consider an approximation to the power function of the Lan-Simon-Halperin test developed by Davis and Hardy (1990). (See also Betensky, 1997.)

### 6.3.2 Futility rules based on predictive probability

As we explained in Section 6.3.1, predictive power tests are derived by averaging the conditional power function with respect to the posterior distribution of the treatment difference given the already observed data. Several authors have indicated that predictive power tests are based on a mixture of Bayesian and frequentist methods and therefore

neither Bayesian nor frequentist statisticians may be satisfied [with them] (Jennison and Turnbull, 1990)

and

the result does not have an acceptable frequentist interpretation and, furthermore, this is not the kind of test a Bayesian would apply (Geisser and Johnson, 1994)

This section introduces an alternative approach to Bayesian futility monitoring known as the *predictive probability* approach. Roughly speaking, predictive probability tests are constructed by replacing the frequentist component of predictive power tests (conditional power function) with a Bayesian one (posterior probability of a positive trial outcome). The predictive probability approach is illustrated below using clinical trials with normally distributed and binary response variables from Case studies 3 and 4.

## Normally distributed endpoints

A positive trial outcome is defined within the Bayesian predictive probability framework in terms of the posterior probability of a clinically important treatment effect rather than statistical significance. To see how this change affects futility testing in clinical applications, consider a two-arm trial comparing an experimental treatment to placebo, and assume that its primary endpoint is a continuous, normally distributed variable. A single interim analysis is performed in the trial at the information fraction  $t$  ( $0 < t < 1$ ).

The trial will be declared positive if we demonstrate that the posterior probability of a clinically important improvement is greater than a prespecified confidence level  $\eta$ , i.e.,

$$P(\mu_1 - \mu_2 > \delta | \text{observed data}) > \eta,$$

where  $\mu_1$  and  $\mu_2$  denote the mean treatment effects in the two trial arms,  $\delta$  is a clinically significant treatment difference, and  $0 < \eta < 1$ . Note that  $\eta$  is typically greater than 0.8, and choosing a larger value of the confidence level reduces the likelihood of a positive trial outcome. The treatment difference  $\delta$  can be either a constant or expressed in terms of the standard deviation. In that case, the introduced criterion becomes a function of the effect size  $(\mu_1 - \mu_2)/\sigma$  (here  $\sigma$  denotes the common standard deviation in the two arms).

In order to predict the probability of a positive trial outcome from the data available at an interim analysis, the data collected before the interim are used to predict future data, which are then combined with the observed data to estimate the posterior probability of a clinically significant treatment difference. Assuming that the mean treatment effects  $\mu_1$  and  $\mu_2$  follow normal priors, a closed-form expression for the predictive probability of observing a clinically significant treatment difference upon termination given the interim test statistic  $Z_t$  was derived in Dmitrienko and Wang (2006). It is interesting to note that, with a uniform prior and  $\delta = 0$ , the predictive power and predictive probability methods become identical when  $\eta = 1 - \alpha$ . For example, a predictive probability test with a 90% confidence level ( $\eta = 0.9$ ) is identical to a predictive power test with a one-sided 0.1 level ( $\alpha = 0.1$ ).

Looking at the case of a small common variance of the prior distributions of  $\mu_1$  and  $\mu_2$ , it is easy to demonstrate that the predictive probability test is asymptotically independent of the test statistic  $Z_t$ . Thus, it turns into a deterministic rule. Let  $\mu_1^*$  and  $\mu_2^*$  denote the prior means of the mean treatment effects  $\mu_1$  and  $\mu_2$ . The predictive probability converges to 1 if  $\mu_1^* - \mu_2^* > \delta$  and to 0 otherwise. This means that the predictive probability test will trigger an early termination of the trial regardless of the size of the treatment difference as long as the selected prior distributions meet the condition  $\mu_1^* - \mu_2^* \leq \delta$ . Due to this property, it is advisable to avoid strong prior distributions when setting up futility rules based on the predictive probability method.

The predictive probability approach in clinical trials with continuous, normally distributed endpoints can be implemented using the `%BayesFutilityCont` macro. This macro also supports predictive power calculations discussed in Section

6.3.1 and can be used to compute predictive power with informative priors. The `%BayesFutilityCont` macro has the following arguments:

- **Data** is the name of the input data set with one record per interim look. The data set must include the following variables:
  - **N1** and **N2** are the numbers of patients in the treatment and placebo arms included in the analysis data set at the current interim look.
  - **Mean1** and **Mean2** are the estimates of the mean treatment effects in the treatment and placebo arms at the current interim look.
  - **SD1** and **SD2** are the sample standard deviations in the treatment and placebo arms at the current interim look.
- **Par** is the name of the single-record input data set with the following variables:
  - **NN1** and **NN2** are the projected numbers of patients in the treatment and placebo arms.
  - **Mu1**, **Mu2** and **Sigma** are the means and common standard deviation of the prior distributions of the mean treatment effects in the treatment and placebo arms.
- **Delta** is the clinically significant difference between the trial arms. This parameter is required by the Bayesian predictive probability method and is ignored by the predictive power method.
- **Eta** is the confidence level of the Bayesian predictive probability method. This parameter is required by the Bayesian predictive probability method and is ignored by the predictive power method.
- **Alpha** is the one-sided Type I error probability of the significance test carried out at the end of the trial. This parameter is required by the predictive power method and is ignored by the Bayesian predictive probability method.
- **Prob** is the name of the data set containing the predictive power and predictive probability at each interim look. This data set is generated by the macro and includes the following variables:
  - **Analysis** is the analysis number.
  - **Fraction** is the fraction of the total sample size.
  - **PredPower** is the computed predictive power (predictive power method).
  - **PredProb** is the computed predictive probability (Bayesian predictive probability method).

## **Futility stopping rules based on predictive probability in Case study 4**

Futility stopping rules based on predictive probability in trials with normally distributed endpoints will be illustrated using Case study 4. Recall from Section 6.3 that a lower value of the HAMA total score indicates a beneficial effect in this trial. Since predictive probability was introduced above under the assumption that a positive treatment difference is desirable, the primary endpoint will be defined as the reduction from baseline, which is a positive quantity.

As was pointed out in Section 6.3.1, an important feature of Bayesian approaches to defining the probability of success in a clinical trial is that the choice of a prior distribution for the treatment difference is likely to impact predictive inferences, especially in small proof-of-concept trials. However, no consensus has been reached in the literature regarding a comprehensive prior selection strategy. For example,

Spiegelhalter, Freedman, and Blackburn (1986) recommended to use uniform prior distributions in clinical applications since they are quickly overwhelmed by the data, and, thus, the associated predictive inferences are driven only by the observed data. Johns and Anderson (1999) also used decision rules based on a non-informative prior. In contrast, Herson (1979) stressed that a uniform prior is often undesirable since it assigns too much weight to extreme values of the treatment difference.

Elicitation of the prior distribution from clinical experts can be a complex and time-consuming activity. See Freedman and Spiegelhalter (1983) for a description of an elaborate interview of 18 physicians to quantify their prior beliefs about the size of the treatment effect in a clinical trial. Alternatively, we can use simple rules of thumb for selecting a prior distribution proposed in the literature. For example, Herson (1979) provided a set of guidelines for the selection of a prior distribution in the binary case. According to the guidelines, a beta prior with the coefficient of variation of 0.1 indicates a high degree of confidence, whereas the value of 0.5 corresponds to low confidence. A similar rule can be adopted to choose priors for the parameters of a normal distribution. For example, considering the clinical trial from Case study 4, Table 6.26 presents two sets of parameters of normal priors that could be considered in this setting. In both cases, the means of the prior distributions are equal to the expected reduction in the HAMA total score in the two trial arms, i.e., 12 in the experimental arm and 8 in the placebo arm. The common standard deviation is chosen in such a way that the coefficient of variation with respect to the average HAMA change is equal to 0.1 or 1. The low confidence scenario defines two prior distributions that are virtually non-informative.

**TABLE 6.26 Parameters of the normal prior for the mean changes in the HAMA score in Case study 4**

Normal prior distribution	Treatment arm		Placebo arm	
	Mean	SD	Mean	SD
“Low confidence” prior	12	10	8	10
“High confidence” prior	12	1	8	1

Program 6.13 invokes the `%BayesFutilityCont` macro to carry out the predictive probability tests at the three interim looks in the generalized anxiety disorder trial. The interim analysis data are included in the `genanx` data, and the `lowconf` data set defines the parameters of the assumed prior distributions of the mean effects in the treatment and placebo arms. This includes the means (`mu1` and `mu2`) and common standard deviation (`sigma`) of the two prior distributions. These parameters correspond to the “low confidence” prior defined in Table 6.26. The `lowconf` data set also includes the projected sample sizes `nn1` and `nn2` of 41 patients per arm at the final analysis. The predictive probability calculations can be easily repeated for the high confidence priors from Table 6.26 or any other set of prior distributions. The two data sets are passed to `%BayesFutilityCont` with three sets of parameters. Specifically, `delta` is set to 1, 2, and 3, and `eta` is set to 0.9. The chosen value of the confidence level  $\eta$  corresponds to a one-sided significance level of 0.1 in the predictive power framework. Note that the `alpha` parameter will be ignored in predictive probability calculations.

#### **PROGRAM 6.13 Predictive probability test in Case study 4**

```
data genanx;
  input n1 mean1 sd1 n2 mean2 sd2;
  datalines;
  10 9.2 7.3 11 8.4 6.4
  20 9.4 6.7 20 8.1 6.9
  31 8.9 7.4 30 9.1 7.7
run;
```

```

data lowconf;
    input nn1 nn2 mu1 mu2 sigma;
    datalines;
        41 41 12 8 10
    run;

%BayesFutilityCont(data=genanx,par=lowconf,delta=1,eta=0.9,
    alpha=0.1,prob=delta1);
%BayesFutilityCont(data=genanx,par=lowconf,delta=2,eta=0.9,
    alpha=0.1,prob=delta2);
%BayesFutilityCont(data=genanx,par=lowconf,delta=3,eta=0.9,
    alpha=0.1,prob=delta3);

data delta1;
    set delta1;
    predprob1=predprob;
    keep analysis predprob1;
run;

data delta2;
    set delta2;
    predprob2=predprob;
    keep analysis predprob2;
run;

data delta3;
    set delta3;
    predprob3=predprob;
    keep analysis predprob3;
run;

data predprob;
    merge delta1 delta2 delta3;
run;

proc print data=predprob noobs;
    var analysis predprob1 predprob2 predprob3;
run;

```

The output of Program 6.13 is summarized in Table 6.27. The results will help us examine the relationship between the magnitude of the clinically significant treatment difference  $\delta$  and predictive probability of a positive final outcome in the generalized anxiety disorder trial. The table shows the predictive probability of observing a 1-, 2-, and 3-point improvement in the mean HAMA change compared to placebo at each of the three interim looks. As expected, the predictive probability declined quickly with the increasing  $\delta$ . Consider, for example, the first interim analysis. The predictive probability of observing a 1-point mean treatment difference at the end of the trial was 21.4%, whereas the predictive probability for 2-and 3-point differences were equal to 11.6% and 5.6%, respectively. These success probabilities were quite low, and it would no longer be worthwhile continuing the trial to its

**TABLE 6.27 Output from Program 6.13**

Interim look	Predictive probability		
	$\delta = 1$	$\delta = 2$	$\delta = 3$
1	0.2135	0.1164	0.0556
2	0.1523	0.0458	0.0094
3	0.0004	0.0000	0.0000

planned end if the trial's sponsor was interested in detecting a substantial amount of improvement over placebo.

## Binary endpoints

The predictive probability framework introduced above is easily extended to the case of binary outcome variables. Consider the severe sepsis trial from Case study 3, and let  $p_1$  and  $p_2$  denote the 28-day survival rates in the treatment and placebo arms. The trial will be declared successful at the final analysis if the probability of a clinically meaningful difference in survival rates exceeds a prespecified confidence level  $\eta$ , i.e.,

$$P(p_1 - p_2 > \delta | \text{observed data}) > \eta,$$

where  $\delta$  denotes a clinically significant treatment difference.

As in the normal case, the posterior probability of  $p_1 - p_2 > \delta$  depends both on the interim and future data. Dmitrienko and Wang (2006) presented a formula for computing the predictive probability as a function of the clinically meaningful treatment difference  $\delta$ , confidence level  $\eta$ , and prior distributions of the event rates  $p_1$  and  $p_2$ . Using this formula, the `%BayesFutilityBin` macro implements the futility testing method based on predictive probability in clinical trials with binary endpoints. It is worth noting that this macro also supports the mixed Bayesian-frequentist approach, i.e., performs predictive power calculations with informative priors, described in Section 6.3.1. The `%BayesFutilityBin` macro has the following arguments:

- **Data** is the name of the input data set with one record per interim look. The data set must include the following variables:
  - **N1** and **N2** are the numbers of patients in the treatment and placebo arms included in the analysis data set at the current interim look.
  - **Count1** and **Count2** are the observed event counts in the treatment and placebo arms at the current interim look.
- **Par** is the name of the single-record input data set with the following variables:
  - **NN1** and **NN2** are the projected numbers of patients in the treatment and placebo arms.
  - **Alpha1** and **Alpha2** are the  $\alpha$  parameters of beta priors for event rates in the treatment and placebo arms.
  - **Beta1** and **Beta2** are the  $\beta$  parameters of beta priors for event rates in the treatment and placebo arms.
- **Delta** is the clinically significant difference between the trial arms. This parameter is required by the Bayesian predictive probability method and is ignored by the predictive power method.
- **Eta** is the confidence level of the Bayesian predictive probability method. This parameter is required by the Bayesian predictive probability method and is ignored by the predictive power method.
- **Alpha** is the one-sided Type I error probability of the significance test carried out at the end of the trial. This parameter is required by the predictive power method and is ignored by the Bayesian predictive probability method.
- **Prob** is the name of the data set containing the predictive power and predictive probability at each interim look. This data set includes the same four variables as

the output data set in the %BayesFutilityCont macro, i.e., `Analysis`, `Fraction`, `PredPower`, and `PredProb`.

### **Futility stopping rules based on predictive probability in Case study 3**

The predictive probability approach to setting up futility stopping rules in trials with binary endpoints will be illustrated using the severe sepsis trial from Case study 3. To perform predictive probability calculations in this setting, prior distributions for the parameters of interest (i.e., 28-day survival rates in the two trial arms) must be defined. Table 6.28 defines two sets of beta distributions corresponding to low and high confidence scenarios. The calculations assumed that the 28-day survival rates in the treatment and placebo arms are equal to 79% and 70%, respectively. The prior distributions were chosen using a version of the Herson rule. As was mentioned previously, Herson (1979) proposed a simple rule to facilitate the selection of prior distributions in clinical trials with a binary response variable:

- “Low confidence” prior: Beta distribution with the mean equal to the expected event rate (e.g., expected 28-day survival rate) and coefficient of variation equal to 0.5.
- “Medium confidence” prior: Beta distribution with the mean equal to the expected event rate and coefficient of variation equal to 0.25.
- “High confidence” prior: Beta distribution with the mean equal to the expected event rate and coefficient of variation equal to 0.1.

Once the mean  $m$  and coefficient of variation  $c$  of a beta distribution have been specified, its parameters are given by

$$\alpha = (1 - m)/c^2 - m, \quad \beta = \alpha(1 - m)/m.$$

In practice, we often need to slightly modify the described rule. First, the Herson rule is applicable to event rates ranging from 0.3 to 0.7. Outside of this range, it might not be possible to achieve the coefficient of variation of 0.5, or even 0.25, with a bell-shaped beta distribution, and U-shaped priors are clearly undesirable in practice. For this reason, uniform priors with  $\alpha = \beta = 1$  in both trial arms were selected to represent a low confidence scenario in the severe sepsis trial. Further, even prior distributions with the coefficient of variation around 0.1 are fairly weak and are completely dominated by the data in large trials. To specify high confidence priors in the severe sepsis trial, we set the coefficient of variation to 0.025.

**TABLE 6.28 Prior distribution of 28-day survival rates in Case study 3**

Beta prior distribution	Treatment arm		Placebo arm	
	$\alpha$	$\beta$	$\alpha$	$\beta$
“Low confidence” prior	1	1	1	1
“High confidence” prior	335	89	479	205

Program 6.14 performs predictive probability calculations at the six interim looks in the severe sepsis trial by using the %BayesFutilityBin macro. The interim analysis data in the severe sepsis trial are defined in the `sevsep` data. The parameters of the beta distributions corresponding to the “low confidence” scenario from Table 6.28 are included in the `lowconf` data set. In addition to the prior distribution parameters, this data set defines the projected sample sizes in each trial arm at the final analysis. Further, suppose that we are interested in computing the predictive

probability of observing any improvement in 28-day survival as well as observing a 5% and 10% absolute increase in the 28-day survival rate at the end of the trial. In this case, the `delta` parameter is set to 0, 0.05, and 0.1. The confidence level of the Bayesian predictive probability test (`Eta` parameter) is equal to 0.9. The calculations are performed for the set of low confidence priors shown in Table 6.28. Lastly, as in Program 6.13, the `alpha` parameter will be ignored in the predictive probability calculations.

#### PROGRAM 6.14 Predictive probability test in Case study 3

```

data sevsep;
    input n1 count1 n2 count2;
    datalines;
        55   33   45   29
        79   51   74   49
        101  65   95   62
        117  75   115  77
        136  88   134  88
        155  99   151  99
    run;

data lowconf;
    input nn1 nn2 alpha1 alpha2 beta1 beta2;
    datalines;
        212 212 1 1 1 1
    run;

%BayesFutilityBin(data=sevsep,par=lowconf,delta=0,eta=0.9,
    alpha=0.1,prob=delta1);
%BayesFutilityBin(data=sevsep,par=lowconf,delta=0.05,eta=0.9,
    alpha=0.1,prob=delta2);
%BayesFutilityBin(data=sevsep,par=lowconf,delta=0.1,eta=0.9,
    alpha=0.1,prob=delta3);

data delta1;
    set delta1;
    predprob1=predprob;
    keep analysis predprob1;
run;

data delta2;
    set delta2;
    predprob2=predprob;
    keep analysis predprob2;
run;

data delta3;
    set delta3;
    predprob3=predprob;
    keep analysis predprob3;
run;

data predprob;
    merge delta1 delta2 delta3;
run;

proc print data=predprob noobs;
    var analysis predprob1 predprob2 predprob3;
run;
```

The predictive probability calculations at the six interim looks in the severe sepsis trial are presented in Table 6.29. This table lists the predictive probability values under three choices of the clinically significant treatment difference  $\delta$  ( $\delta = 0, 0.05$  and  $0.1$ ). Recall from the beginning of this section that, for normally distributed outcomes, the predictive probability test is closely related to the predictive power test when  $\eta = 1 - \alpha$  and the clinically meaningful treatment difference  $\delta$  is set to 0. A similar relationship is observed in the binary case as well. If  $\delta = 0$ , the predictive probability test will roughly correspond to the predictive power test with a one-sided significance level of 0.1.

Table 6.29 demonstrates that, as expected, larger values of  $\delta$  were associated with lower predictive probabilities of a positive trial outcome. For example, if the trial's sponsor would like to demonstrate with a 90% confidence that the experimental treatment improves the 28-day survival rate by 5% or 10% (in absolute terms), the predictive probability of a positive trial outcome at the very first interim look drops to 4.8% and 1.3%, respectively. It appears prudent to suspend the patient enrollment at the very first look in this trial since the negative conclusion is unlikely to change even if the trial continued to the planned end.

**TABLE 6.29 Output from Program 6.14**

Interim look	Predictive probability		
	$\delta = 0$	$\delta = 0.05$	$\delta = 0.1$
1	0.1264	0.0481	0.0131
2	0.1059	0.0279	0.0040
3	0.0809	0.0144	0.0010
4	0.0204	0.0013	0.0000
5	0.0188	0.0007	0.0000
6	0.0087	0.0001	0.0000

### 6.3.3 Other applications of stochastic curtailment methods

Sections 6.3.1 and 6.3.2 introduced three families of stochastic curtailment tests and focused mostly on their use in futility monitoring. In what follows, we will briefly describe other applications of stochastic curtailment methods in late-stage clinical trials.

Stochastic curtailment tests have found numerous applications in adaptive trials where the trial's design can be modified based on the findings observed at one or more interim analyses. The following decision rules are commonly used in adaptive trials:

- Sample size adjustment rules support an option to increase the total sample size or target number of events at an interim analysis.
- Population selection rules are used in clinical trials with two or more predefined patient populations, e.g., the overall population of patients with a condition of interest and a subpopulation defined using baseline variables. The trial's design might be modified at an interim analysis by restricting patient enrollment to a subpopulation.
- Treatment selection rules are conceptually similar to population selection rules in that they support a decision to discontinue patient enrollment in one or more trial arms based on the data available at an interim analysis.

Sample size adjustment rules are most commonly used in adaptive trials compared to other decision rules. The goal of sample size adjustment is to improve the probability of success in a trial if the treatment difference computed at an interim

analysis appears to be lower than anticipated. A stochastic curtailment method is applied to predict the number of patients that would be required to achieve the desired level of the success probability at the final analysis. If the predicted sample size is greater than the predefined sample size, a decision is made to increase the total number of patients to be enrolled in the trial up to a reasonable cap. This general approach goes back to Proschan and Hunsberger (1995) and has been used in numerous clinical trials. While sample size adjustment rules are typically formulated using conditional power, data-driven design adaptations can also be performed using Bayesian predictive methods, see, for example, Wang (2007).

Adaptive clinical trials with population selection rules are becoming increasingly popular, mostly due to the interest in precision medicine or tailored therapies. An oncology Phase III trial presented in Brannath et al. (2009) serves as a good example of adaptive population selection trials. This trial was conducted to evaluate the effect of a novel treatment on progression-free survival compared to control in the overall trial population as well as a predefined subpopulation of biomarker-positive patients. The subpopulation was selected because the patients with a biomarker-positive status were expected to experience enhanced treatment benefit compared to biomarker-negative patients. A Bayesian stochastic curtailment method was applied to compute the predictive probabilities of success in the overall population and biomarker-positive subpopulation at the interim analysis. The calculations were to support the decision to continue enrolling all eligible patients after the interim look or to discontinue patient enrollment in the complementary subpopulation of biomarker-negative patients if the novel treatment did not appear to be effective in this subpopulation.

Finally, as stated above, adaptive rules can be aimed at selecting the best dose or doses of the experimental treatment at an interim look. One of the first trials with adaptive dose selection was the indacaterol trial in patients with chronic obstructive pulmonary disease (Barnes et al., 2010). The trial used a two-stage design, and four doses of indacaterol, along with positive and negative controls, were studied in the first stage. An interim analysis was conducted at the end of the first stage to evaluate the preliminary efficacy results. The dose selection rules used in this particular trial relied on the observed treatment differences, but stochastic curtailment methods are often applied in similar settings to predict the probability of success in each dosing arm at the planned end of the trial. This information can be used to select the most effective doses to be studied in the second stage of the trial, and the dosing arms with a low predicted probability of success can be dropped.

A critical component of the adaptive designs described above is ensuring that the overall Type I error rate is protected at a desired level. Data-driven design adaptations are known to affect the Type I error rate, and several methods have been developed to provide Type I error rate control in adaptive designs. This includes the combination function and conditional rejection probability approaches. See Wassmer and Brannath (2016) for more information on these approaches.

Another important application of stochastic curtailment methods deals with probability of success predictions in Phase III development programs based on the results of one or more Phase II clinical trials conducted in the same patient population. The predictions are typically performed within a Bayesian framework and rely on predictive power and predictive probability calculations. See, for example, Chuang-Stein (2006), Chuang-Stein et al. (2011), Wang et al., (2013) and Wang (2015). Specifically, the sponsor of a Phase III program can treat the end of Phase II development as an interim analysis and evaluate the probability of success in the Phase III trials using the predictive power or predictive probability approaches. By taking advantage of these Bayesian strategies, the sponsor can efficiently account for the variability around the treatment effects estimated from the Phase II trials. This leads to more reliable predictions that will ultimately support the sponsor's portfolio-related decisions. For a summary of recent developments in this general area, see Patra et al. (2017).

### 6.3.4 Summary

This section reviewed three approaches to setting up stochastic curtailment tests in clinical trials: frequentist conditional power methods, mixed Bayesian-frequentist methods based on predictive power, and Bayesian predictive probability methods. As their name suggests, stochastic curtailment tests rely on stochastic arguments in order to decide when to curtail sampling. Specifically, a clinical trial is terminated at an interim analysis if the probability of success at the trial's end predicted from the interim analysis data is unacceptably low.

The three families of tests differ in how they define the predicted probability of success at the scheduled end of the trial:

- **Conditional power tests** evaluate the probability of observing a statistically significant treatment difference at the planned end of the trial given the interim data. Conditional power tests include the Lan-Simon-Halperin test that relies heavily on the hypothesized value of the treatment effect as well as other tests that are based on a sample estimate of the treatment difference to predict the distribution of the future data. The latter are generally more appealing in clinical trial applications than tests tied to an assumed value of the treatment difference. The conditional power methods can be implemented using PROC SEQTEST.
- **Predictive power tests** incorporate more data-driven elements into the decision rule. This is achieved by averaging the conditional power function with respect to a posterior distribution of the treatment difference. The clinical trial is terminated as soon as the average conditional power (known as the *predictive power*) falls below a certain threshold. An assumed prior distribution of the treatment effect can have great influence on the resulting futility rule. For example, selecting a non-informative prior that is quickly dominated by the data increases the chances of early stopping due to futility. PROC SEQTEST currently supports predictive power tests with non-informative priors. If informative priors are available, predictive power calculations in clinical trials with normally distributed endpoints and binary endpoints can be performed using the %BayesFutilityCont and %BayesFutilityBin macros, respectively.
- **Predictive probability tests** provide a fully Bayesian solution to the futility testing problem. Within the predictive probability framework, a positive trial outcome is defined in terms of the posterior probability of a clinically important treatment effect rather than statistical significance. Under a non-informative prior, predictive probability tests yield results similar to those produced by predictive power tests and turn into deterministic rules when the variance of the prior distribution is small. For this reason, it is advisable to avoid strong priors when setting up futility rules based on the predictive probability method. Futility stopping rules based on Bayesian predictive probability tests can be implemented using the %BayesFutilityCont and %BayesFutilityBin macros written by the authors.

Although the Type II error rate of trial designs with futility monitoring is inflated, which leads to power loss, adjustments to account for this potential inflation are rarely performed in practice. This is due mainly to the fact that stochastic curtailment tests minimize the probability of early stopping compared to repeated significance tests described in Section 6.2.

Finally, due to their simplicity, futility tests reviewed in this section might be more appealing in practice than computationally complex futility monitoring strategies based on error spending functions. In addition, the futility boundary of a group sequential design depends heavily on the assumed value of the treatment difference that might no longer be supported by the data. Therefore, the repeated significance

approach tends to be less robust than the stochastic curtailment approach (for example, Bayesian methods described in this section).

## 6.4 References

---

- Armitage, P., McPherson, C.K., Rowe, B.C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A* 132, 235-244.
- Barnes, P.J., Pocock, S.J., Magnussen, H., Iqbal, A., Kramer, B., Higgins, M., Lawrence, D. (2010). Integrating indacaterol dose selection in a clinical study in COPD using an adaptive seamless design. *Pulmonary Pharmacology and Therapeutics* 23, 165-171.
- Bauer, P., Bretz, F., Dragalin, V., König, F., Wassmer, G. (2016). 25 years of confirmatory adaptive designs: Opportunities and pitfalls. *Statistics in Medicine* 35, 325-347.
- Betensky, R.A. (1997). Early stopping to accept  $H_0$  based on conditional power: approximations and comparisons. *Biometrics* 53, 794-806.
- Bernardo, M.V.P., Ibrahim, J.G. (2000). Group sequential designs for cure rate models with early stopping in favour of the null hypothesis. *Statistics in Medicine* 19, 3023-3035.
- Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy on oncology. *Statistics in Medicine* 28, 1445-1463.
- Chang, W.H., Chuang-Stein, C. (2004). Type I error and power in trials with one interim futility analysis. *Pharmaceutical Statistics* 3, 51-59.
- Chuang-Stein, C. (2006). Sample size and the probability of a successful trial. *Pharmaceutical Statistics* 5, 305-309.
- Chuang-Stein, C., Kirby, S., French, J., Kowalski, K., Marshall, S., Smith, M.K., Bycott, P., Beltangady, M. (2011). A quantitative approach for making Go/No-Go decisions in drug development. *Pharmaceutical Statistics* 45, 187-202.
- Danielson, L., Carlier, J., Burzykowski, T., Buyse, M. (2014). Implementation issues in adaptive design trials. *Practical Considerations for Adaptive Trial Design and Implementation*. He, W., Pinheiro, J., Kuznetsova, O. (editors). New York: Springer.
- Davis, B.R., Hardy, R.J. (1990). Upper bounds for Type I and type II error rates in conditional power calculations. *Communications in Statistics. Part A* 19, 3571-3584.
- DeMets, D.L., Ware, J.H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika* 67, 651-660.
- DeMets, D.L., Lan, K.K.G. (1984). An overview of sequential methods and their application in clinical trials. *Communications in Statistics. Part A* 13, 2315-2338.
- Dmitrienko, A., Govindarajulu, Z. (2000). Sequential confidence regions for maximum likelihood estimates. *Annals of Statistics* 28, 1472-1501.
- Dmitrienko, A., Wang, M.D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine* 25, 2178-2195.
- Dmitrienko, A., Paux, G., Pulkstenis, E., Zhang, J. (2016). Tradeoff-based optimization criteria in clinical trials with multiple objectives and adaptive designs. *Journal of Biopharmaceutical Statistics* 26, 120-140.

- Dmitrienko, A., Pulkstenis, E. (editors). (2017). *Clinical Trial Optimization Using R*. Boca Raton, FL: Chapman and Hall/CRC Press.
- Ellenberg, S.S., Fleming, T.R., DeMets, D.L. (2002). *Data Monitoring Committees In Clinical Trials: A Practical Perspective*. New York: Wiley.
- EMA (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. European Medicines Agency.
- Emerson, S.S., Fleming, T.R. (1989). Symmetric group sequential test designs. *Biometrics* 45, 905-923.
- Enas G.G., Dornseif, B.E., Sampson C.B., Rockhold F.W., Wu, J. (1989). Monitoring versus interim analysis of clinical trials: Perspective from the pharmaceutical industry. *Controlled Clinical Trials* 10, 57-70.
- FDA (2010). Draft guidance for industry: Adaptive design clinical trials for drugs and biologics. U.S. Food and Drug Administration.
- FDA (2016). Guidance for industry and Food and Drug Administration staff: Adaptive designs for medical device clinical studies. U.S. Food and Drug Administration.
- Freedman, L.S., Spiegelhalter, D.J. (1983). The assessment of the subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician* 32, 153-160.
- Gallo, P., DeMets, D., LaVange, L. (2014). Considerations for interim analyses in adaptive trials, and perspectives on the use of DMCs. *Practical Considerations for Adaptive Trial Design and Implementation*. He, W., Pinheiro, J., Kuznetsova, O. (editors). Springer, New York.
- Geisser, S. (1992). On the curtailment of sampling. *Canadian Journal of Statistics* 20, 297-309.
- Geisser, S., Johnson, W. (1994). Interim analysis for normally distributed observables. In *Multivariate analysis and Its Applications. IMS Lecture Notes* 263-279.
- Halperin, M., Lan, K.K.G., Ware, J.H., Johnson, N.J., DeMets, D.L. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials* 3, 311-323.
- Herson, J. (1979). Predictive probability early termination plans for Phase II clinical trials. *Biometrics* 35, 775-783.
- 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.
- Jennison, C., Turnbull, B.W. (1989). Interim analysis: the repeated confidence interval approach. *Journal of the Royal Statistical Society. Series B* 51, 305-361.
- Jennison, C., Turnbull, B.W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science* 5, 299-317.
- Jennison, C., Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL:Chapman and Hall/CRC Press.
- Johns, D., Anderson, J.S. (1999). Use of predictive probabilities in Pahse II and Phase III clinical trials. *Journal of Biopharmaceutical Statistics* 9, 67-79.
- 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.
- Lan, K.K.G., Simon, R., Halperin, M. (1982). Stochastically curtailed tests in long-term clinical terms. *Communications in Statistics. Sequential Analysis* 1, 207-219.

- Lan, K.K.G., DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70, 659-663.
- Lan, K.K.G., DeMets, D.L. (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics* 45, 1017-1020.
- Lan, K.K.G., Wittes, J. (1988). The B-Value: A tool for monitoring data. *Biometrics* 44, 579-585.
- Lan, K.K.G., Lachin, J.M., Bautista, O.M. (2003). Over-ruling a group sequential boundary---a stopping rule versus a guideline. *Statistics in Medicine* 22, 3347-3355.
- Lee, J.J., Liu, D.D. (2008). A predictive probability design for Phase II cancer clinical trials. *Clinical Trials* 5, 93-106.
- O'Brien, P.C., Fleming, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* 35, 549-556.
- O'Brien, P.C. (1990). Group sequential methods in clinical trials. *Statistical Methodology in the Pharmaceutical Sciences* 291-311.
- Pampallona, S., Tsiatis, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. *Journal of Statistical Planning and Inference* 42, 19-35.
- Pampallona, S., Tsiatis, A.A., Kim, K. (2001). Interim monitoring of group sequential trials using spending functions for the Type I and Type II error probabilities. *Drug Information Journal* 35, 1113-1121.
- Patra, K., Wang, M.D., Zhang, J., Dane, A., Metcalfe, P., Frewer, P., Pulkstenis, E. (2017). *Clinical Trial Optimization Using R*. Dmitrienko, A., Pulkstenis, E. (editors). Boca Raton, FL: Chapman and Hall/CRC Press.
- Pepe, M.S., Anderson, G.L. (1992). Two-stage experimental designs: Early stopping with a negative result. *Applied Statistics* 41, 181-190.
- Pocock, S.J. (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64, 191-199.
- Proschan, M.A., Follman, D.A., Waclawiw, M.A. (1992). Effects of assumption violations on Type I error rate in group sequential monitoring. *Biometrics* 48, 1131-1143.
- Proschan, M., Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics* 51, 1315-1324.
- Proschan, M.A., Lan, K.K.G., Wittes, J.T. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. New York: Springer.
- Siegmund, D. (1978). Estimation following sequential tests. *Biometrika* 65, 341-349.
- Spiegelhalter, D.J., Freedman, L.S., Blackburn, P.R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials* 7, 8-17.
- Tsiatis, A.A., Rosner, G.L., Mehta, C.R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* 40, 797-803.
- Van Den Berghe, G., Wouters, P., Weekers, F., Verwaest, C., Bruyninckx, F., Schetz, I., Vlasselaers, D., Ferdinande, P., Lauwers, P., Bouillon, R. (2001). Intensive insulin therapy in critically ill patients. *The New England Journal of Medicine*. 345, 1359-1367.
- Wang, M.D. (2007). Sample size re-estimation by Bayesian prediction. *Biometrical Journal* 49, 365-377.
- Wang, M.D. (2015). Applications of probability of study success in clinical drug development. *Applied Statistics in Biomedicine and Clinical Trials Design*. Chen, Z., Liu, A., Qu, Y., Tang, L., Ting, N., Tsong, Y. (editors). Springer, New York.

- Wang, S.K., Tsiatis, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 43, 193-199.
- Wang, Y., Fu, H., Kulkarni, P., Kaiser, C. (2013). Evaluating and utilizing probability of study success in clinical development. *Clinical Trials* 10, 407-413.
- Ware, J.H., Muller, J.E., Braunwald, E. (1985). The futility index: An approach to the cost-effective termination of randomized clinical trials. *American Journal of Medicine*. 78, 635-643.
- Wassmer, G., Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. New York: Springer.
- Whitehead, J., Stratton, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics* 39, 227-236.
- Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* 73, 573-581.
- Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Second edition. Chichester, UK: Wiley.
- Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statistics in Medicine* 18, 2271-2286.

# Chapter 7

## Analysis of Incomplete Data

Geert Molenberghs (Universiteit Hasselt and KU Leuven)

Michael G. Kenward

7.1	Introduction	319
7.2	Case Study	322
7.3	Data Setting and Methodology	324
7.4	Simple Methods and MCAR	334
7.5	Ignorable Likelihood (Direct Likelihood)	338
7.6	Direct Bayesian Analysis (Ignorable Bayesian Analysis)	341
7.7	Weighted Generalized Estimating Equations	344
7.8	Multiple Imputation	349
7.9	An Overview of Sensitivity Analysis	362
7.10	Sensitivity Analysis Using Local Influence	363
7.11	Sensitivity Analysis Based on Multiple Imputation and Pattern-Mixture Models	371
7.12	Concluding Remarks	378
7.13	References	378

More often than not, empirical studies are prone to incompleteness. For about half a century, methods have been developed to address this issue in data analysis. Older methods are relatively simple to use, but their validity is rightly called into question. With increasing computational power and software tools available, more flexible methods have come within reach. This chapter sketches a general taxonomy (Rubin, 1976) within which incomplete data methods can be placed. It then focuses on broadly valid methods that can be implemented within the SAS environment, thereby commenting on their relative advantages and disadvantages. All methods are illustrated using real data, and sufficiently generic SAS code is offered. Both Gaussian and non-Gaussian outcomes are given treatment. Apart from standard analysis tools, sensitivity analysis to examine the impact of non-verifiable model assumptions is addressed.

### 7.1 Introduction

---

In a longitudinal study, each unit is measured on several occasions. It is not unusual in practice for some sequences of measurements to terminate early for reasons outside the control of the investigator, and any unit so affected is called a dropout. It might, therefore, be necessary to accommodate dropout in the modeling process.

Early work on missing values was largely concerned with algorithmic and computational solutions to the induced lack of balance or deviations from the intended

study design (Afifi and Elashoff, 1966; Hartley and Hocking, 1971). This was followed by the development of general algorithms such as expectation-maximization (EM) (Dempster, Laird, and Rubin, 1977), and data imputation and augmentation procedures (Rubin 1987). These methods, combined with contemporary powerful computing resources and the progressive implementation of advanced methods in the SAS system, have addressed the problem in important ways. There is also the very difficult and important question of assessing the impact of missing data on subsequent statistical inference. This has received attention in particular in the setting of clinical trials (Little et al., 2010). Several authors give practical advice regarding the use of incomplete data methods (Mallinckrodt, 2013; O'Kelly and Ratitch, 2014), while others focus on the broad methodological underpinnings (Molenberghs and Kenward, 2007), or on specific methods, such as multiple imputation (MI; van Buuren, 2012; Carpenter and Kenward, 2013). The edited volumes by Fitzmaurice et al. (2009) and Molenberghs et al. (2015) present overviews of the longitudinal data and incomplete data state of research, respectively.

When referring to the missing-value, or non-response, process, we will use terminology of Little and Rubin (2014, Chapter 6). A non-response process is said to be *missing completely at random* (MCAR) if missingness is independent of both unobserved and observed data and *missing at random* (MAR) if, conditional on the observed data, missingness is independent of the unobserved measurements. A process that is neither MCAR nor MAR is termed *non-random* (MNAR). In the context of likelihood or Bayesian inferences, when the parameters describing the measurement process are functionally independent of the parameters describing the missingness process, and provided some mild regularity conditions hold, MCAR and MAR are *ignorable*, while a non-random process is non-ignorable. In the same vein, MI is valid under MAR. The method offers an attractive Monte Carlo-based alternative to direct likelihood and Bayesian inferences. For frequentist inferences, only a strong MCAR assumption is a sufficient condition for ignorability. This is relevant when discussing such methods as *generalized estimating equations* (GEE; Liang and Zeger, 1986).

We will pay particular attention to these methods, because of their relevance and the ease with which they can be implemented in SAS, thanks to the availability of a suite of SAS procedures. This implies that historic methods such as *complete case analysis* (CC) and *last observation carried forward* (LOCF) will be de-emphasized, in line with Little et al. (2010). Indeed, valid inference can be obtained through a likelihood-based analysis, a Bayesian analysis, or multiple imputation, without the need for modeling the dropout or missingness process. Likelihood-based analyses of longitudinal data can easily be conducted *without additional data manipulation* using, for example, the SAS procedures MIXED, GLIMMIX, NLMIXED, or related procedures (Verbeke and Molenberghs, 2000), without additional complication or effort. Thanks to the availability and flexibility of the procedures MI and MIANALYZE, multiple imputation is also rather straightforward to conduct. Furthermore, whereas a proper GEE analysis (i.e., valid under MAR) requires substantial additional programming with PROC GENMOD, the newer GEE procedure has made so-called *weighted GEE* (WGEE) particularly easy.

At the same time, we cannot avoid reflecting on the status of MNAR-based approaches. In realistic settings, the reasons for missingness or dropout are varied and hard to know with sufficient certainty. It is, therefore, difficult to fully justify on *a priori* grounds the assumption of MAR. At first sight, this calls for a further shift towards MNAR models. However, careful considerations have to be made, the most important of which is that no modeling approach, whether MAR or MNAR, can recover the lack of information that occurs due to incompleteness of the data.

First, under MAR, a standard analysis would follow, if we would be entirely sure of the MAR nature of the mechanism. However, it is only rarely the case that such an assumption is known to hold (Murray and Findlay, 1988). Nevertheless, ignorable

analyses may provide reasonably stable results, even when the assumption of MAR is violated, in the sense that such analyses constrain the behavior of the unseen data to be similar to that of the observed data (Mallinckrodt et al., 2001ab). A discussion of this phenomenon in the survey context can be found in Rubin, Stern, and Vehovar (1995). These authors argue that, in well conducted experiments (some surveys and many confirmatory clinical trials), the assumption of MAR is often to be regarded as a realistic one. Second, and very important for confirmatory trials, an MAR analysis can be specified *a priori* without additional work relative to a situation with complete data. Third, while MNAR models are more general and explicitly incorporate the dropout mechanism, the inferences they produce are typically highly dependent on untestable and often implicit assumptions built in regarding the distribution of the unobserved measurements given the observed ones. The quality of the fit to the observed data need not reflect at all the appropriateness of the implied structure governing the unobserved data. This point is irrespective of the MNAR route taken, whether a parametric model of the type of Diggle and Kenward (1994) is chosen, or a semi-parametric approach such as in Robins, Rotnitzky, and Scharfstein (1998). Hence in any incomplete-data setting there cannot be anything like a definitive analysis.

Thus, arguably, in the presence of MNAR missingness, a wholly satisfactory analysis of the data is not feasible. In fact, modeling in this context often rests on strong (untestable) assumptions and relatively little evidence from the data themselves. Glynn, Laird, and Rubin (1986) indicated that this is typical for selection models. It is somewhat less the case for pattern-mixture models (Little 1993, 1994; Hogan and Laird 1997), although caution should be used (Thijs, Molenberghs, and Verbeke, 2000). This awareness and the resulting skepticism about fitting MNAR models initiated the search for methods to investigate the results with respect to model assumptions and for methods allowing to assess influences in the parameters describing the measurement process, as well as the parameters describing the non-random part of the dropout mechanism. Several authors have suggested various types of sensitivity analyses to address this issue (Molenberghs, Kenward, and Goetghebeur, 2001; Scharfstein, Rotnitzky, and Robins, 1999; Van Steen et al., 2001; and Verbeke et al., 2001). Verbeke et al. (2001) and Thijs, Molenberghs, and Verbeke (2000) developed a local influence-based approach for the detection of subjects that strongly influence the conclusions. These authors focused on the Diggle and Kenward (1994) model for continuous outcomes. Van Steen et al. (2001) adapted these ideas to the model of Molenberghs, Kenward and Lesaffre (1997), for monotone repeated ordinal data. Jansen et al. (2003) focused on the model family proposed by Baker, Rosenberger, and DerSimonian (1992). Recently, considerable research attention has been devoted to the use of pattern-mixture models, combined with multiple imputation, as a viable route for sensitivity analysis (Carpenter and Kenward, 2013; Carpenter, Roger, and Kenward, 2013). In summary, to explore the impact of deviations from the MAR assumption on the conclusions, we should ideally conduct a sensitivity analysis, within which MNAR models can play a major role.

The rest of the chapter is organized as follows. The clinical trial that will be used throughout the chapter is introduced in Section 7.2. The general datasetting is introduced in Section 7.3, as well as a formal framework for incomplete longitudinal data. A brief overview on the problems associated with simple methods is presented in Section 7.4. In subsequent sections, key methods are examined: ignorable likelihood (Section 7.5); ignorable Bayesian analysis (Section 7.6); generalized estimating equations (Section 7.7); and multiple imputation (Section 7.8). A brief introduction to sensitivity analysis is given in Section 7.9. Generally sensitivity analysis tools are discussed in Section 7.10, while in Section 7.11 we focus on sensitivity analysis tools that make use of multiple imputation.

The SAS code and data sets included in this chapter are available on the book's website at <http://support.sas.com/publishing/authors/dmitrienko.html>.

## 7.2 Case Study

---

### **EXAMPLE: Age-related macular degeneration trial**

These data arise from a randomized multi-center clinical trial comparing an experimental treatment (interferon- $\alpha$ ) to a corresponding placebo in the treatment of patients with age-related macular degeneration. In this book, we focus on the comparison between placebo and the highest dose (6 million units daily) of interferon- $\alpha$  ( $Z$ ). But the full results of this trial have been reported elsewhere (Pharmacological Therapy for Macular Degeneration Study Group 1997). Patients with macular degeneration progressively lose vision. In the trial, the patients' visual acuity was assessed at different time points (4 weeks, 12 weeks, 24 weeks, and 52 weeks) through their ability to read lines of letters on standardized vision charts. These charts display lines of 5 letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). The raw patient's visual acuity is the total number of letters correctly read. In addition, we often refer to each line with at least 4 letters correctly read as a "line of vision." The primary endpoint of the trial was the loss of at least 3 lines of vision at 1 year, compared to their baseline performance (a binary endpoint). The secondary endpoint of the trial was the visual acuity at 1 year (treated as a continuous endpoint). Buyse and Molenberghs (1998) examined whether the patient's performance at 6 months could be used as a surrogate for their performance at 1 year with respect to the effect of interferon- $\alpha$ . They looked at whether the loss of 2 lines of vision at 6 months could be used as a surrogate for the loss of at least 3 lines of vision at 1 year (Table 7.1). They also looked at whether visual acuity at 6 months could be used as a surrogate for visual acuity at 1 year.

**TABLE 7.1** **The Age-related Macular Degeneration Trial. Loss of at least 3 lines of vision at 1 year according to loss of at least 2 lines of vision at 6 months and according to randomized treatment group (placebo versus interferon- $\alpha$ ).**

		12 months	
		Placebo	Active
6 months		0	1
No event (0)		56	9
Event (1)		8	30
		31	9
		9	38

Table 7.2 shows the visual acuity (mean and standard error) by treatment group at baseline, at 6 months, and at 1 year.

**TABLE 7.2** **The Age-related Macular Degeneration Trial. Mean (standard error) of visual acuity at baseline, at 6 months and at 1 year according to randomized treatment group (placebo versus interferon- $\alpha$ ).**

Time point	Placebo	Active	Total
Baseline	55.3 (1.4)	54.6 (1.3)	55.0 (1.0)
6 months	49.3 (1.8)	45.5 (1.8)	47.5 (1.3)
1 year	44.4 (1.8)	39.1 (1.9)	42.0 (1.3)

Visual acuity can be measured in several ways. First, we can record the number of letters read. Alternatively, dichotomized versions (at least 3 lines of vision lost) can

be used as well. Therefore, these data will be useful to illustrate methods for the joint modeling of continuous and binary outcomes, with or without taking the longitudinal nature into account. In addition, though there are 190 subjects with both month 6 and month 12 measurements available, the total number of longitudinal profiles is 240, but for only 188 of these have the four follow-up measurements been made.

Thus, indeed, 50 incomplete subjects could be considered for analysis as well. Both intermittent missingness as well as dropout occurs. An overview is given in Table 7.3.

**TABLE 7.3** The Age-related Macular Degeneration Trial. Overview of missingness patterns and the frequencies with which they occur. 'O' indicates observed and 'M' indicates missing.

Measurement occasion				Number	%
4 wks	12 wks	24 wks	52 wks		
Completers					
O	O	O	O	188	78.33
Dropouts					
O	O	O	M	24	10.00
O	O	M	M	8	3.33
O	M	M	M	6	2.50
M	M	M	M	6	2.50
Non-monotone missingness					
O	O	M	O	4	1.67
O	M	M	O	1	0.42
M	O	O	O	2	0.83
M	O	M	M	1	0.42

Thus, 78.33% of the profiles are complete, while 18.33% exhibit monotone missingness. Out of the latter group, 2.5% or 6 subjects have no follow-up measurements. The remaining 3.33%, representing 8 subjects, have intermittent missing values. Thus, as in many of the examples seen already, dropout dominates intermediate patterns as the source of missing data.

<b>Age-related Macular Degeneration Trial. Partial printout.</b>	CRF	TRT	VISUAL0	VISUAL4	VISUAL12	VISUAL24	VISUAL52	lesion
	1002	4	59	55	45	.	.	3
	1003	4	65	70	65	65	55	1
	1006	1	40	40	37	17	.	4
	1007	1	67	64	64	64	68	2
	1010	4	70	.	.	.	.	1
	1110	4	59	53	52	53	42	3
	1111	1	64	68	74	72	65	1
	1112	1	39	37	43	37	37	3
	1115	4	59	58	49	54	58	2
	1803	1	49	51	71	71	.	1
	1805	4	58	50	.	.	.	1
	...							

The original outcome (number of letters correctly read on a vision chart or its difference with the baseline reading) can be considered continuous for practical purposes. The derived dichotomous outcome (defined as number of letters read, has increased versus decreased when compared with baseline) will be considered as well.

Note that of the 52 subjects with incomplete follow-up, 8 exhibit a non-monotone pattern. While this will not hamper direct-likelihood analyses, Bayesian analyses, or multiple imputation, it is a challenge for weighted GEE and will need to be addressed.

## 7.3 Data Setting and Methodology

---

Assume that for subject  $i = 1, \dots, N$  in the study a sequence of responses,  $Y_{ij}$  is designed to be measured at occasions  $j = 1, \dots, n_i$ . The outcomes are grouped into a vector of random variables  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})'$ . In addition, define a dropout indicator  $D_i$  for the occasion at which dropout occurs and make the convention that  $D_i = n_i + 1$  for a complete sequence. It is often handy to split the vector  $\mathbf{Y}_i$  into observed ( $\mathbf{Y}_i^o$ ) and missing ( $\mathbf{Y}_i^m$ ) components, respectively. Dropout is a particular case of monotone missingness. To have a monotone pattern, there has to exist a permutation of the components of  $\mathbf{Y}_i$  for all  $i$  simultaneously, such that, if a component is missing, then all later components are missing as well. For example, consider a vector of length four:  $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ , with all but the second component  $Y_{i2}$  fully observed. Then the ordering  $(Y_{i1}, Y_{i3}, Y_{i4}, Y_{i2})'$  satisfies the definition of monotone missingness. A counterexample is when, for every  $i$ , either  $Y_{i1}$  or  $Y_{i2}$  is observed. In that case, no monotone re-ordering is possible. For this definition to be meaningful, we need to have a balanced design in the sense of a common set of measurement occasions across all study subjects. Other patterns are referred to as non-monotone or intermittent missingness.

In principle, we would like to consider the density of the full data  $f(\mathbf{y}_i, d_i | \boldsymbol{\theta}, \boldsymbol{\psi})$ , where the parameter vectors  $\boldsymbol{\theta}$  and  $\boldsymbol{\psi}$  describe the measurement and missingness processes, respectively. Covariates are assumed to be measured but, for notational simplicity, suppressed from notation unless strictly needed.

The taxonomy, constructed by Rubin (1976), further developed in Little and Rubin (1987, with later editions in 2002 and 2014) and informally sketched in Section 7.1, is based on the factorization

$$f(\mathbf{y}_i, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i | \boldsymbol{\theta}) f(d_i | \mathbf{y}_i, \boldsymbol{\psi}), \quad (7.3.1)$$

where the first factor is the marginal density of the measurement process, and the second one is the density of the missingness process, conditional on the outcomes. Factorization (7.3.1) forms the basis of *selection modeling* as the second factor corresponds to the (self-)selection of individuals into “observed” and “missing” groups. An alternative taxonomy can be built based on so-called *pattern-mixture models* (Little, 1993, 1994). These are based on the factorization

$$f(\mathbf{y}_i, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}_i | d_i, \boldsymbol{\theta}) f(d_i | \boldsymbol{\psi}). \quad (7.3.2)$$

Indeed, (7.3.2) can be seen as a mixture of different populations, characterized by the observed pattern of missingness.

In the selection modeling framework, let us first describe a measurement and missingness model in turn, and then formally introduce and comment on ignorability.

### 7.3.1 Linear Mixed Models

Assume that we want to perform a longitudinal analysis of a continuous outcome. We then often assume a linear mixed-effects model, sometimes with an additional serial correlation:

$$\mathbf{Y}_i = X_i \boldsymbol{\beta} + Z_i \mathbf{b}_i + \boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i}, \quad (7.3.3)$$

(Verbeke and Molenberghs, 2000) where  $\mathbf{Y}_i$  is the  $n$ -dimensional response vector for subject  $i$ ,  $1 \leq i \leq N$ ;  $N$  is the number of subjects;  $X_i$  and  $Z_i$  are  $(n_i \times p)$  and  $(n_i \times q)$  known design matrices;  $\beta$  is the  $p$  dimensional vector containing the fixed effects; and  $\mathbf{b}_i \sim N(\mathbf{0}, G)$  is the  $q$  dimensional vector containing the random effects. The residual components  $\varepsilon_i$  are decomposed as  $\varepsilon_i = \varepsilon_{(1)i} + \varepsilon_{(2)i}$  in which  $\varepsilon_{(2)i}$  is a component of serial correlation and  $\varepsilon_{(1)i} \sim N(\mathbf{0}, \sigma^2 I_{n_i})$  is an extra component of measurement error. Thus, serial correlation is captured by the realization of a Gaussian stochastic process,  $\varepsilon_{(2)i}$ , which is assumed to follow a  $N(\mathbf{0}, \tau^2 H_i)$  law. The serial covariance matrix  $H_i$  only depends on  $i$  through the number  $n$  of observations and through the time points  $t_{ij}$  at which measurements are taken. The structure of the matrix  $H_i$  is determined through the autocorrelation function  $\rho(t_{ij} - t_{ik})$ . This function decreases such that  $\rho(0) = 1$  and  $\rho(+\infty) = 0$ . Further,  $G$  is a general  $(q \times q)$  covariance matrix with  $(i, j)$  element  $d_{ij} = d_{ji}$ . Finally,  $\mathbf{b}_1, \dots, \mathbf{b}_N, \varepsilon_{(1)1}, \dots, \varepsilon_{(2)N}, \varepsilon_{(2)1}, \dots, \varepsilon_{(2)N}$  are assumed to be independent. Inference is based on the marginal distribution of the response  $\mathbf{Y}_i$  which, after integrating over random effects, can be expressed as

$$\mathbf{Y}_i \sim N(X_i \beta, Z_i G Z_i' + \Sigma_i). \quad (7.3.4)$$

Here,  $\Sigma_i = \sigma^2 I_{n_i} + \tau^2 H_i$  is a  $(n \times n)$  covariance matrix that groups the measurement error and serial components. Further, we define  $V_i = Z_i G Z_i' + \Sigma_i$  as the general covariance matrix of  $\mathbf{Y}_i$ .

The most commonly used SAS procedure to fit linear mixed models is PROC MIXED. The fixed-effect structure is specified via the MODEL statement, while the random-effects structure is entered using the RANDOM statement. If, in addition, serial correlation is assumed to be present, the REPEATED statement can be added. Also, several marginal models derived from a linear mixed-effects model can be specified directly using the REPEATED statement. For details, we refer to Verbeke and Molenberghs (2000).

### 7.3.2 Generalized Linear Mixed Models

Perhaps the most commonly encountered subject-specific (or random-effects) model for arbitrary outcome data type is the generalized linear mixed model (GLMM). A general framework for mixed-effects models can be expressed as follows.

It is assumed that, conditionally on  $q$ -dimensional random effects  $\mathbf{b}_i$  that are drawn independently from  $N(\mathbf{0}, G)$ , the outcomes  $Y_{ij}$  are independent with densities of the form

$$f_i(y_{ij} | \mathbf{b}_i, \beta, \phi) = \exp \left\{ \phi^{-1}[y_{ij}\theta_{ij} - \psi(\theta_{ij})] + c(y_{ij}, \phi) \right\},$$

with  $\eta(\mu_{ij}) = \eta(E(Y_{ij} | \mathbf{b}_i)) = \mathbf{x}'_{ij} \beta + \mathbf{z}'_{ij} \mathbf{b}_i$  for a known link function  $\eta(\cdot)$ , with  $\mathbf{x}_{ij}$  and  $\mathbf{z}_{ij}$   $p$ -dimensional and  $q$ -dimensional vectors of known covariate values; with  $\beta$  a  $p$ -dimensional vector of unknown fixed regression coefficients; with  $\phi$  a scale parameter; and with  $\theta_{ij}$  the natural (or canonical) parameter. Further, let  $f(\mathbf{b}_i | G)$  be the density of the  $N(\mathbf{0}, G)$  distribution for the random effects  $\mathbf{b}_i$ .

Due to the above independence assumption, this model is often referred to as a *conditional independence* model. This assumption is the basis of the implementation in the NLINMIXED procedure. Just as in the linear mixed model case, the model can be extended with residual correlation, in addition to the one induced by the random effects. Such an extension can be implemented in the SAS procedure GLIMMIX, and its predecessor the GLIMMIX macro. It is relevant to realize that GLIMMIX can be used without random effects as well, thus effectively producing a marginal model, with estimates and standard errors similar to the ones obtained with GEE (see Section 7.3.4).

In general, unless a fully Bayesian approach is followed, inference is based on the marginal model for  $\mathbf{Y}_i$  which is obtained from integrating out the random effects. The likelihood contribution of subject  $i$  then becomes

$$f_i(\mathbf{y}_i|\boldsymbol{\beta}, G, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|G) d\mathbf{b}_i$$

from which the likelihood for  $\boldsymbol{\beta}$ ,  $D$ , and  $\phi$  is derived as

$$\begin{aligned} L(\boldsymbol{\beta}, G, \phi) &= \prod_{i=1}^N f_i(\mathbf{y}_i|\boldsymbol{\beta}, G, \phi) \\ &= \prod_{i=1}^N \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|G) d\mathbf{b}_i. \end{aligned} \quad (7.3.5)$$

The key problem in maximizing the obtained likelihood is the presence of  $N$  integrals over the  $q$ -dimensional random effects. In some special cases, these integrals can be worked out analytically. However, since no analytic expressions are available for these integrals, numerical approximations are needed. Here, we will focus on the most frequently used methods to do so. In general, the numerical approximations can be subdivided into those that are based on the approximation of the integrand; those based on an approximation of the data; and those that are based on the approximation of the integral itself. An extensive overview of a number of available approximations can be found in Tuerlinckx et al. (2004), Pinheiro and Bates (2000), and Skrondal and Rabe-Hesketh (2004). Finally, to simplify notation, it will be assumed that natural link functions are used, but straightforward extensions can be applied.

When integrands are approximated, the goal is to obtain a tractable integral such that closed-form expressions can be obtained, making the numerical maximization of the approximated likelihood feasible. Several methods have been proposed, but basically all come down to Laplace-type approximations of the function to be integrated (Tierney and Kadane 1986).

A second class of approaches is based on a decomposition of the data into the mean and an appropriate error term, with a Taylor series expansion of the mean, which is a nonlinear function of the linear predictor. All methods in this class differ in the order of the Taylor approximation and/or the point around which the approximation is expanded. More specifically, we consider the decomposition

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} = h(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i) + \varepsilon_{ij}, \quad (7.3.6)$$

in which  $h(\cdot)$  equals the inverse link function  $\eta^{-1}(\cdot)$ , and where the error terms have the appropriate distribution with variance equal to  $\text{Var}(Y_{ij}|\mathbf{b}_i) = \phi v(\mu_{ij})$  for  $v(\cdot)$ , the usual variance function in the exponential family. Note that, with the natural link function,

$$v(\mu_{ij}) = \frac{\partial h}{\partial \eta}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i).$$

Several approximations of the mean  $\mu_{ij}$  in (7.3.6) can be considered. One possibility is to derive a linear Taylor expansion of (7.3.6) around current estimates  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{b}}_i$  of the fixed effects and random effects, respectively. This will result in the expression

$$\mathbf{Y}_i^* \equiv \widehat{W}_i^{-1}(\mathbf{Y}_i - \hat{\mu}_i) + X_i \hat{\boldsymbol{\beta}} + Z_i \hat{\mathbf{b}}_i \approx X_i \boldsymbol{\beta} + Z_i \mathbf{b}_i + \varepsilon_i^*, \quad (7.3.7)$$

with  $\widehat{W}_i$  equal to the diagonal matrix with diagonal entries equal to  $v(\widehat{\mu}_{ij})$ , and for  $\varepsilon_i^*$  equal to  $\widehat{W}_i^{-1}\varepsilon_i$ , which still has mean zero. Note that (7.3.7) can be viewed as a

linear mixed model for the pseudo data  $\mathbf{Y}_i^*$ , with fixed effects  $\beta$ , random effects  $\mathbf{b}_i$ , and error terms  $\varepsilon_i^*$ .

This immediately yields an algorithm for fitting the original generalized linear mixed model. Given starting values for the parameters  $\beta$ ,  $G$ , and  $\phi$  in the marginal likelihood, empirical Bayes estimates are calculated for  $\mathbf{b}_i$ , and pseudo data  $\mathbf{Y}_i^*$  are computed. Then, the approximate linear mixed model (7.3.7) is fitted, yielding updated estimates for  $\beta$ ,  $G$ , and  $\phi$ . These are then used to update the pseudo data, and this whole scheme is iterated until convergence is reached.

The resulting estimates are called *penalized quasi-likelihood* estimates (PQL) in the literature (e.g., Molenberghs and Verbeke, 2005), or *pseudo-quasi-likelihood* in the documentation of the GLIMMIX procedure because they can be obtained from optimizing a quasi-likelihood function that only involves first and second-order conditional moments, augmented with a penalty term on the random effects. The pseudo-likelihood terminology derives from the fact that the estimates are obtained by (restricted) maximum likelihood of the pseudo-response or working variable.

An alternative approximation is very similar to the PQL method, but is based on a linear Taylor expansion of the mean  $\mu_{ij}$  in (7.3.6) around the current estimates  $\hat{\beta}$  for the fixed effects and around  $\mathbf{b}_i = \mathbf{0}$  for the random effects. The resulting estimates are called *marginal quasi-likelihood* estimates (MQL). We refer to Breslow and Clayton (1993) and Wolfinger and O'Connell (1993) for more details. Since the linearizations in the PQL and the MQL methods lead to linear mixed models, the implementation of these procedures is often based on feeding updated pseudo data into software for the fitting of linear mixed models. However, it should be emphasized that the results from these fittings, which are often reported intermediately, should be interpreted with great care. For example, reported (log)likelihood values correspond to the assumed normal model for the pseudo data and should not be confused with (log-)likelihood for the generalized linear mixed model for the actual data at hand. Further, fitting of linear mixed models can be based on maximum likelihood (ML) as well as restricted maximum likelihood (REML) estimation. Hence, within the PQL and MQL frameworks, both methods can be used for the fitting of the linear model to the pseudo data, yielding (slightly) different results. Finally, the quasi-likelihood methods discussed here are very similar to the method of linearization for fitting generalized estimating equations (GEE). The difference is that here, the correlation between repeated measurements is modelled through the inclusion of random effects, conditionally on which repeated measures are assumed independent. But in the GEE approach, this association is modelled through a marginal working correlation matrix.

Note that, when there are no random effects, both this method and GEE reduce to a marginal model, the difference being in the way that the correlation parameters are estimated. In both cases, it is possible to allow for misspecification of the association structure by resorting to empirically corrected standard errors. When this is done, the methods are valid under MCAR. In case we would have confidence in the specified correlation structure, purely model-based inference can be conducted, and, hence, the methods are valid when missing data are MAR.

A third method of numerical approximation is based on the approximation of the integral itself. Especially in cases where the above two approximation methods fail, this numerical integration turns out to be very useful. Of course, a wide toolkit of numerical integration tools, available from the optimization literature, can be applied. Several of those have been implemented in various software tools for generalized linear mixed models. A general class of quadrature rules selects a set of abscissas and constructs a weighted sum of function evaluations over those. In the particular context of random-effects models, so-called *adaptive* quadrature rules can be used (Pinheiro and Bates, 1995, 2000), where the numerical integration is centered around the EB estimates of the random effects. The number of quadrature points is then selected in terms of the desired accuracy.

To illustrate the main ideas, we consider Gaussian and adaptive Gaussian quadrature, designed for the approximation of integrals of the form  $\int f(z)\phi(z)dz$ , for an known function  $f(z)$  and for  $\phi(z)$  the density of the (multivariate) standard normal distribution. We will first standardize the random effects such that they get the identity covariance matrix. Let  $\delta_i$  be equal to  $\delta_i = G^{-1/2}\mathbf{b}_i$ . We then have that  $\delta_i$  is normally distributed with mean  $\mathbf{0}$  and covariance  $I$ . The linear predictor then becomes  $\theta_{ij} = \mathbf{x}'_{ij}\beta + z'_{ij}G^{1/2}\delta_i$ , so the variance components in  $G$  have been moved to the linear predictor. The likelihood contribution for subject  $i$ , expressed in the original parameters, then equals

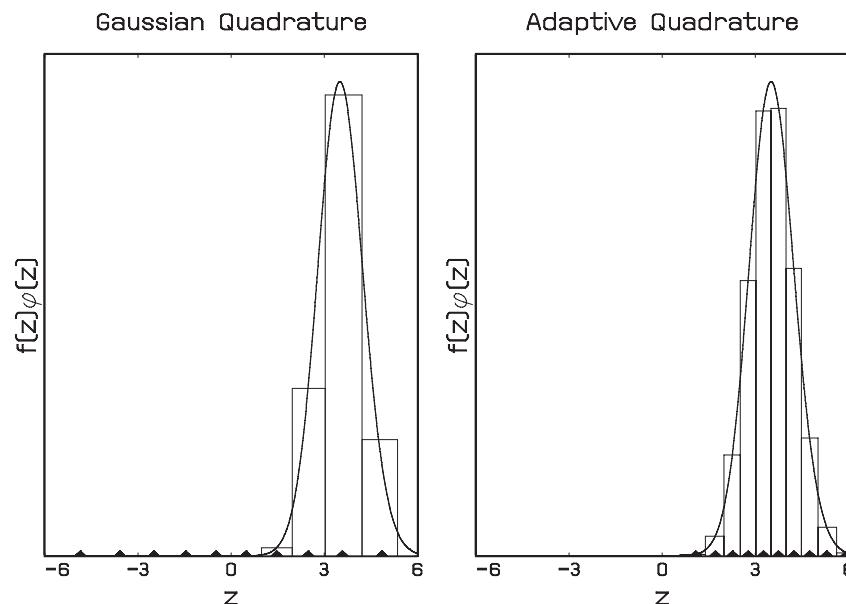
$$f_i(\mathbf{y}_i|\boldsymbol{\beta}, G, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\mathbf{b}_i, \boldsymbol{\beta}, \phi) f(\mathbf{b}_i|G) d\mathbf{b}_i. \quad (7.3.8)$$

Obviously, (7.3.8) is of the form  $\int f(z)\phi(z)dz$  as required to apply (adaptive) Gaussian quadrature.

In Gaussian quadrature,  $\int f(z)\phi(z)dz$  is approximated by the weighted sum

$$\int f(z)\phi(z)dz \approx \sum_{q=1}^Q w_q f(z_q).$$

$Q$  is the order of the approximation. The higher  $Q$ , the more accurate the approximation will be. Further, the so-called nodes (or quadrature points)  $z_q$  are solutions to the  $Q$ th order Hermite polynomial, while the  $w_q$  are well-chosen weights. The nodes  $z_q$  and weights  $w_q$  are reported in tables. Alternatively, an algorithm is available for calculating all  $z_q$  and  $w_q$  for any value  $Q$  (Press et al., 1992). In case of univariate integration, the approximation consists of subdividing the integration region in intervals, and approximating the surface under the integrand by the sum of surfaces of the so-obtained approximating rectangles. An example is given in the left window of Figure 7.1, for the case of  $Q = 10$  quadrature points. A similar interpretation is possible for the approximation of multivariate integrals. Note that the figure immediately highlights one of the main disadvantages of (non-adaptive) Gaussian quadrature, i.e., the fact that the quadrature points  $z_q$  are chosen based on  $\phi(z)$ , independent of the function  $f(z)$  in the integrand. Depending on the support of



**Figure 7.1**  
Graphical illustration of Gaussian (left window) and adaptive Gaussian (right window) quadrature of order  $Q = 10$ . The black triangles indicate the position of the quadrature points, while the rectangles indicate the contribution of each point to the integral.

$f(z)$ , the  $z_q$  will or will not lie in the region of interest. Indeed, the quadrature points are selected to perform well in case  $f(z)\phi(z)$  approximately behaves like  $\phi(z)$ , i.e., like a standard normal density function. This will be the case, for example, if  $f(z)$  is a polynomial of a sufficiently low order. In our applications, however, the function  $f(z)$  will take the form of a density from the exponential family---hence, an exponential function. It might then be helpful to re-scale and shift the quadrature points such that more quadrature points lie in the region of interest. This is shown in the right window of Figure 7.1, and is called adaptive Gaussian quadrature.

In general, the higher the order  $Q$ , the better the approximation will be of the  $N$  integrals in the likelihood. Typically, adaptive Gaussian quadrature needs (many) fewer quadrature points than classical Gaussian quadrature. On the other hand, adaptive Gaussian quadrature requires for each unit the numerical maximization of a function of the form  $\ln(f(z)\phi(z))$  for the calculation of  $\hat{z}$ . This implies that adaptive Gaussian quadrature is much more time consuming.

Since fitting of GLMMs is based on maximum likelihood principles, inferences for the parameters are readily obtained from classical maximum likelihood theory.

The Laplace method (Tierny and Kadane, 1986) has been designed to approximate integrals of the form

$$I = \int e^{Q(\mathbf{b})} d\mathbf{b}, \quad (7.3.9)$$

where  $Q(\mathbf{b})$  is a known, unimodal, and bounded function of a  $q$ -dimensional variable  $\mathbf{b}$ . Let  $\hat{\mathbf{b}}$  be the value of  $\mathbf{b}$  for which  $Q$  is maximized. We then have that the second-order Taylor expansion of  $Q(\mathbf{b})$ , which is of the form

$$Q(\mathbf{b}) \approx Q(\hat{\mathbf{b}}) + \frac{1}{2}(\mathbf{b} - \hat{\mathbf{b}})'Q''(\hat{\mathbf{b}})(\mathbf{b} - \hat{\mathbf{b}}), \quad (7.3.10)$$

for  $Q''(\hat{\mathbf{b}})$  equal to the Hessian of  $Q$ , i.e., the matrix of second-order derivative of  $Q$ , evaluated at  $\hat{\mathbf{b}}$ . Replacing  $Q(\mathbf{b})$  in (7.3.9) by its approximation in (7.3.10), we obtain

$$I \approx (2\pi)^{q/2} \left| -Q''(\hat{\mathbf{b}}) \right|^{-1/2} e^{Q(\hat{\mathbf{b}})}.$$

Clearly, each integral in (7.3.5) is proportional to an integral of the form (7.3.9), for functions  $Q(\mathbf{b})$  given by

$$Q(\mathbf{b}) = \phi^{-1} \sum_{j=1}^{n_i} [y_{ij}(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}) - \psi(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b})] - \frac{1}{2}\mathbf{b}'D^{-1}\mathbf{b},$$

such that Laplace's method can be applied here. Note that the mode  $\hat{\mathbf{b}}$  of  $Q$  depends on the unknown parameters  $\boldsymbol{\beta}$ ,  $\phi$ , and  $D$ , such that in each iteration of the numerical maximization of the likelihood,  $\hat{\mathbf{b}}$  will be re-calculated conditionally on the current values for the estimates for these parameters.

The Laplace approximation is exact when  $Q(\mathbf{b})$  is a quadratic function of  $\mathbf{b}$ , i.e., if the integrands in (7.3.5) are exactly equal to normal kernels. Interpreting these integrands as unnormalized posterior distributions of the random effects  $\mathbf{b}_i$ , it is known from the Bayesian literature (Gelman et al., 1995) that this will be the case only in very special examples such as linear models, or provided that the number  $n_i$  of repeated measurements for all subjects are sufficiently large.

To fit GLMMs, the SAS procedures GLIMMIX and NLMIXED are obvious choices. While a variety of GLMMs can be fitted using both procedures, there are fundamental differences. GLIMMIX is restricted to generalized linear mixed models, whereas NLMIXED allows for fully nonlinear (mixed) models. For this

reason, GLIMMIX models are specified in a conventional, symbolic way (e.g., using syntax of the form  $Y=X1\ X2\ X1*X2$ ), whereas in NLMIXED the user programs the mean and, where appropriate, variance functions, including fixed and random effects. GLIMMIX allows for serial correlation, using a REPEATED statement. Both procedures allow for multiple RANDOM statements. The integration options in GLIMMIX include PQL, MQL, Laplace approximation, and adaptive Gaussian quadrature. The corresponding options for NLMIXED are adaptive and non-adaptive Gaussian quadrature. Both allow for a variety of updating algorithms and tuning parameters. Because GLIMMIX is restricted to GLMM and, hence, can efficiently make use of generalized linear model features (exponential family results, the use of linear predictors, etc.), it is generally faster and stabler when both procedures can be used. However, NLMIXED offers additional flexibility thanks to the open programming abilities.

Using NLMIXED, the conditional distribution of the data, given the random effects, is specified in the MODEL statement. Valid distributions are:

- $\text{normal}(m, v)$ : Normal with mean  $m$  and variance  $v$ ,
- $\text{binary}(p)$ : Bernoulli with probability  $p$ ,
- $\text{binomial}(n, p)$ : Binomial with count  $n$  and probability  $p$ ,
- $\text{gamma}(a, b)$ : Gamma with shape  $a$  and scale  $b$ ,
- $\text{negbin}(n, p)$ : Negative binomial with count  $n$  and probability  $p$ ,
- $\text{poisson}(m)$ : Poisson with mean  $m$ ,
- $\text{general}(\ell\ell)$ : General model with log-likelihood  $\ell\ell$ .

The general structure is especially convenient when a non-conventional model is fitted. The RANDOM statement defines the random effects and their distribution. The procedure requires the data to be ordered by subject.

The valid distributions for the GLIMMIX procedure are: beta, binary, binomial, exponential, gamma, Gaussian (normal), geometric, inverse Gaussian, lognormal, multinomial, negative binomial, Poisson, and central  $t$ . Each one of them has a default link function attached to them. Users have the ability to deviate from these, but should check whether an alternative choice is coherent with the natural range of the outcome type. For example, a probit link instead of a logit link is also a sensible choice for binary outcomes, while a log link would usually be problematic for interval-type data.

### 7.3.3 Likelihood-based Approaches

Consider, for the sake of argument, a continuous longitudinal outcome. Assume that incompleteness is due to dropout only, and that the first measurement  $Y_{i1}$  is obtained for everyone. The model for the dropout process can be based on, for example, a logistic regression for the probability of dropout at occasion  $j$ , given the subject is still in the study. We denote this probability by  $g(\mathbf{h}_{ij}, y_{ij})$  in which  $\mathbf{h}_{ij}$  is a vector containing all responses observed up to but not including occasion  $j$ , as well as relevant covariates. We then assume that  $g(\mathbf{h}_{ij}, y_{ij})$  satisfies

$$\text{logit}[g(\mathbf{h}_{ij}, y_{ij})] = \text{logit}[\text{pr}(D_i = j | D_i \geq j, \mathbf{y}_i)] = \mathbf{h}_{ij}\boldsymbol{\psi} + \omega y_{ij}, \quad (7.3.11)$$

$i = 1, \dots, N$ . When  $\omega$  equals zero, the posited dropout model is MAR, and all parameters can be estimated easily using SAS since the measurement model for which we use a linear mixed model and the dropout model, assumed to follow a logistic regression, can then be fitted separately. If  $\omega \neq 0$ , the posited dropout process is MNAR. Model (7.3.11) provides the building blocks for the dropout

process  $f(d_i|\mathbf{y}_i, \psi)$ . As a cautionary note, we should not lose sight of the fact that the true nature of the dropout mechanism cannot be determined based on observed data alone (Molenberghs et al., 2008), pointing to the need for sensitivity analysis.

Rubin (1976) and Little and Rubin (2014) have shown that, under MAR and mild regularity conditions (parameters  $\theta$  and  $\psi$  are functionally independent), likelihood-based and Bayesian inferences are valid when the missing data mechanism is ignored (see also Verbeke and Molenberghs, 2000). Practically speaking, the likelihood of interest is then based upon the factor  $f(\mathbf{y}_i^o|\theta)$ . This is called *ignorability*. We return to this in more detail in Sections 7.5 and 7.6.

The practical implication is that a software module with likelihood estimation facilities and with the ability to handle incompletely observed subjects manipulates the correct likelihood, providing valid parameter estimates and likelihood ratio values. A few cautionary remarks are in place. First, when at least part of the scientific interest is directed towards the nonresponse process, obviously both processes need to be considered. Still, under MAR, both processes can be modeled and parameters estimated separately. Second, likelihood inference is often surrounded with references to the sampling distribution (e.g., to construct precision estimators and for statistical hypothesis tests; Kenward and Molenberghs, 1998). However, the practical implication is that standard errors and associated tests, when based on the observed rather than the expected information matrix and given that the parametric assumptions are correct, are valid. Third, it may be hard to fully rule out the operation of an MNAR mechanism. This point was brought up in the introduction and will be discussed further in Sections 7.9–7.10. Fourth, a full longitudinal analysis is necessary, even when interest lies, for example, in a comparison between the two treatment groups at the last occasion. In the latter case, the fitted model can be used as the basis for inference at the last occasion. A common criticism is that a model needs to be considered. However, it should be noted that, in many clinical trial settings, the repeated measures are balanced in the sense that a common (and often limited) set of measurement times is considered for all subjects, allowing the a priori specification of a saturated model (e.g., full group by time interaction model for the fixed effects and unstructured variance-covariance matrix). Such an ignorable linear mixed model specification is given in Mallinckrodt et al. (2001ab).

### 7.3.4 Generalized Estimating Equations

#### Overview

Two sometimes quoted issues with full likelihood approaches are the computational complexity they entail and their vulnerability to model assumptions. When we are mainly interested in first-order marginal mean parameters and pairwise association parameters, i.e., second-order moments, a full likelihood procedure can be replaced by quasi-likelihood methods (McCullagh and Nelder, 1989). In quasi-likelihood, the mean response is expressed as a parametric function of covariates; and the variance is assumed to be a function of the mean up to possibly unknown scale parameters. Wedderburn (1974) first noted that likelihood and quasi-likelihood theories coincide for exponential families and that the quasi-likelihood “estimating equations” provide consistent estimates of the regression parameters  $\beta$  in any generalized linear model, even for choices of link and variance functions that do not correspond to exponential families.

For clustered and repeated data, Liang and Zeger (1986) proposed so-called *generalized estimating equations* (GEE or GEE1), which require only the correct specification of the univariate marginal distributions provided we are willing to adopt “working” assumptions about the association structure. They estimate the parameters associated with the expected value of an individual’s vector of binary

responses and phrase the working assumptions about the association between pairs of outcomes in terms of marginal correlations. The method combines estimating equations for the regression parameters  $\beta$  with moment-based estimating for the correlation parameters entering the working assumptions.

Prentice (1988) extended their results to allow joint estimation of probabilities and pairwise correlations. Lipsitz, Laird, and Harrington (1991) modified the estimating equations of Prentice (1988) to allow modeling of the association through marginal odds ratios rather than marginal correlations. When adopting GEE1 we do not use information of the association structure to estimate the main effect parameters. As a result, it can be shown that GEE1 yields consistent main effect estimators, even when the association structure is misspecified. However, severe misspecification can seriously affect the efficiency of the GEE1 estimators. In addition, GEE1 should be avoided when some scientific interest is placed on the association parameters.

A second-order extension of these estimating equations (GEE2) that include the marginal pairwise association as well has been studied by Liang, Zeger, and Qaqish (1992). They note that GEE2 is nearly fully efficient though bias might occur in the estimation of the main effect parameters when the association structure is misspecified.

Carey, Zeger, and Diggle (1993) proposed so-called *alternating logistic regressions* (ALR), applicable to repeated binary data with logit link and the association modeled using odds ratios. See also Molenberghs and Verbeke (2005). While they allow for association-modeling, they are computationally simpler than GEE2.

### Some Methodological Detail

After this short overview of the GEE approach, the GEE methodology will now be explained a little further. We start by recalling the score equations, to be solved when computing maximum likelihood estimates under a marginal normal model  $\mathbf{y}_i \sim N(X_i\beta, V_i)$ :

$$\sum_{i=1}^N X'_i (A_i^{1/2} R_i A_i^{1/2})^{-1} (\mathbf{y}_i - X_i \beta) = \mathbf{0}, \quad (7.3.12)$$

in which the marginal covariance matrix  $V_i$  has been decomposed in the form  $V_i = A_i^{1/2} R_i A_i^{1/2}$ , with  $A_i$  the diagonal matrix; with the marginal variances along the main diagonal; and with  $R_i$  equal to the marginal correlation matrix. Second, the score equations to be solved when computing maximum likelihood estimates under a marginal generalized linear model, assuming independence of the responses within units (i.e., ignoring the repeated measures structure), are given by:

$$\sum_{i=1}^N \frac{\partial \mu_i}{\partial \beta'} (A_i^{1/2} I_{n_i} A_i^{1/2})^{-1} (\mathbf{y}_i - \mu_i) = \mathbf{0}. \quad (7.3.13)$$

Note that (7.3.12) is of the form (7.3.13) but with the correlations between repeated measures taken into account. A straightforward extension of (7.3.13) that accounts for the correlation structure is

$$S(\beta) = \sum_{i=1}^N \frac{\partial \mu_i}{\partial \beta'} (A_i^{1/2} R_i A_i^{1/2})^{-1} (\mathbf{y}_i - \mu_i) = \mathbf{0}, \quad (7.3.14)$$

which is obtained from replacing the identity matrix  $I_{n_i}$  by a correlation matrix  $R_i = R_i(\alpha)$ , often referred to as the *working* correlation matrix. Usually, the marginal covariance matrix  $V_i = A_i^{1/2} R_i A_i^{1/2}$  contains a vector  $\alpha$  of unknown parameters--leading to  $V_i(\beta, \alpha) = A_i^{1/2}(\beta) R_i(\alpha) A_i^{1/2}(\beta)$ --which is replaced for practical purposes by a consistent estimate.

Assuming that the marginal mean  $\mu_i$  has been correctly specified as  $h(\mu_i) = X_i\beta$ , it can be shown that, under mild regularity conditions, the estimator  $\hat{\beta}$  obtained from solving (7.3.14) is asymptotically normally distributed with mean  $\beta$  and with covariance matrix

$$I_0^{-1} I_1 I_0^{-1}, \quad (7.3.15)$$

where

$$I_0 = \left( \sum_{i=1}^N \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta'} \right), \quad I_1 = \left( \sum_{i=1}^N \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \text{Var}(\mathbf{y}_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta'} \right).$$

In practice,  $\text{Var}(\mathbf{y}_i)$  in (7.3.15) is replaced by  $(\mathbf{y}_i - \mu_i)(\mathbf{y}_i - \mu_i)'$ , which is unbiased on the sole condition that the mean was again correctly specified.

Note that valid inferences can now be obtained for the mean structure, only assuming that the model assumptions with respect to the first-order moments are correct. Note also that, although arising from a likelihood approach, the GEE equations in (7.3.14) cannot be interpreted as score equations corresponding to some full likelihood for the data vector  $\mathbf{y}_i$ .

Liang and Zeger (1986) proposed moment-based estimates for the working correlation. To this end, first define deviations:

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$

and decompose the variance slightly more generally as above in the following way:

$$V_i = \phi A_i^{1/2} R_i A_i^{1/2},$$

where  $\phi$  is an overdispersion parameter.

Some of the more popular choices for the working correlations are independence ( $\text{Corr}(Y_{ij}, Y_{ik}) = 0$ ,  $j \neq k$ ); exchangeability ( $\text{Corr}(Y_{ij}, Y_{ik}) = \alpha$ ,  $j \neq k$ ); AR(1) ( $\text{Corr}(Y_{ij}, Y_{i,j+t}) = \alpha^t$ ,  $t = 0, 1, \dots, n_i - j$ ); and unstructured ( $\text{Corr}(Y_{ij}, Y_{ik}) = \alpha_{jk}$ ,  $j \neq k$ ). Typically, moment-based estimation methods are used to estimate these parameters, as part of an integrated iterative estimation procedure (Aerts, Geys, Molenberghs, and Ryan, 2002). The overdispersion parameter is approached in a similar fashion. The standard iterative procedure to fit GEE, based on Liang and Zeger (1986), is then as follows: (1) compute initial estimates for  $\beta$ , using a univariate GLM (i.e., assuming independence); (2) compute the quantities needed in the estimating equation, such as means and variances; (3) compute Pearson residuals  $e_{ij}$ ; (4) compute estimates for  $\alpha$ ; (5) compute  $R_i(\alpha)$ ; (6) compute an estimate for  $\phi$ ; (7) compute  $V_i(\beta, \alpha) = \phi A_i^{1/2}(\beta) R_i(\alpha) A_i^{1/2}(\beta)$ ; and (8) update the estimate for  $\beta$ :

$$\beta^{(t+1)} = \beta^{(t)} - \left[ \sum_{i=1}^N \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta} \right]^{-1} \left[ \sum_{i=1}^N \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} (\mathbf{y}_i - \mu_i) \right].$$

Steps (2)--(8) are iterated until convergence.

In SAS, three procedures can be used for GEE. First, there is the GENMOD procedure. In its basic form, it fits generalized linear models to univariate data. Adding the REPEATED statement, repeated measures can be analyzed using GEE or ALR, with a suite of working correlation structures available. As of SAS 9.4, the GEE procedure is available. It is essentially a “synonym” to GENMOD for standard GEE, but its main attraction lies in the use of weighted GEE, for which we refer to Section 7.7. As mentioned earlier, the GLIMMIX procedure can also be used, provided no random effects are included, but merely “serial correlation,” using the RANDOM \_residual\_ / syntax, combined with the use of empirically corrected standard errors using the empirical option in the PROC GLIMMIX statement.

## 7.4 Simple Methods and MCAR

---

We will briefly review a number of relatively simple methods that have been and are still in extensive use. For a number of them, MCAR is required, while for others, such as LOCF, the conditions for validity are wholly different. A detailed account is given in Verbeke and Molenberghs (1997, 2000) and Molenberghs and Kenward (2007). The case of clinical trials received specific attention in Molenberghs et al. (2003). The focus will be on the complete case method, where data are removed, on the one hand, and on imputation strategies and where data are filled in on the other hand. Regarding imputation, we distinguish between single and multiple imputation. In the first case, a single value is substituted for every “hole” in the data set, and the resulting data set is analyzed as if it represented the true complete data. Multiple imputation properly acknowledges the uncertainty stemming from filling in missing values rather than observing them (Rubin, 1987; Schafer, 1997), and is deferred to Section 7.8. LOCF will be discussed within the context of imputation strategies, although not every author classifies the method as belonging to the imputation family.

### 7.4.1 Complete Case Analysis

A complete case analysis includes only those cases for analysis for which all  $n_i$  planned measurements were actually recorded. This method has obvious advantages. It is very simple to describe, and, since the data structure is as would have resulted from a complete experiment, standard statistical software can be used. Further, since the complete estimation is done on the same subset of completers, there is a common basis for inference, unlike with the available case methods.

Unfortunately, the method suffers from severe drawbacks. First, there is nearly always a substantial loss of information. For example, suppose there are 20 measurements, with 10% of missing data on each measurement. Suppose further that missingness on the different measurements is independent. Then, the estimated percentage of incomplete observations is as high as 87%. The impact on precision and power is dramatic. Even though the reduction of the number of complete cases will be less dramatic in realistic settings where the missingness indicators  $R_i$  are correlated, the effect just sketched will often undermine a lot of complete case analyses. In addition, severe bias can result when the missingness mechanism is MAR but not MCAR. Indeed, should an estimator be consistent in the complete data problem, then the derived complete case analysis is consistent only if the missingness process is MCAR. Unfortunately, the MCAR assumption is much more restrictive than the MAR assumption.

#### Complete Case Analysis and SAS

The only step required to perform a complete case analysis is deletion of subjects for which not all designed measurements have been obtained. When the data are organized “horizontally,” i.e., one record per subject, this is particularly easy. With “vertically” organized data, slightly more data manipulation is needed, and the SAS macro, discussed below, can be used.

For example, for the age related macular degeneration trial, running the next statement produces the complete case CC data set, for the continuous outcome (‘diff’ is the difference of number of letters correctly read versus, baseline’):

#### PROGRAM 7.1 Preparing the data for complete case analysis (continuous outcome)

```
%cc(data=armd155,id=subject,time=time,response=diff,out=armdcc2);
```

and for the binary outcome ('bindif' is a discretization of 'diff', with 1 for nonnegative values and 0 otherwise). See Program 7.2.

### PROGRAM 7.2 Preparing the data for complete case analysis (discrete outcome)

```
%cc(data=armd111,id=subject,time=time,response=bindif,out=armdcc);
```

Clearly, the CC macro requires four arguments. The **data=** argument is the data set to be analyzed. If not specified, the most recent data set is used. The name of the variable in the data set that contains the identification variable is specified by **id=**, and **time=** specifies the variable indicating the time ordering within a subject. The outcome variable is passed on by means of the **response=** argument, and the name of the output data set, created with the macro, is defined through **out=**.

After performing this data preprocessing, a complete case analysis follows of any type requested by the user, including, but not limited to, longitudinal analysis.

The macro requires records, corresponding to missing values, to be present in the data set. Otherwise, it is assumed that a measurement occasion not included is missing by design.

Upon creation of the new data set, the code for Model (7.5.16), to be presented in Section 7.5 on ignorable likelihood, is given by Program 7.3.

### PROGRAM 7.3 Complete case analysis (continuous outcome)

```
proc mixed data=armdcc2 method=ml;
title 'CC - continuous';
class time treat subject;
model diff = time treat*time / noint solution ddfm=kr;
repeated time / subject=subject type=un;
run;
```

When, in contrast, GEE of the form (7.7.28) is applied to the completers, the following code can be used for standard GEE. See Program 7.4.

### PROGRAM 7.4 Complete case analysis (binary outcome, GEE, PROC GENMOD)

```
proc genmod data=armdcc;
title 'CC - GEE';
class time treat subject;
model bindif = time treat*time / noint dist=binomial;
repeated subject=subject / withinsubject=time type=exch modelse;
run;
```

or see Program 7.5.

### PROGRAM 7.5 Complete case analysis (binary outcome, GEE, PROC GEE)

```
proc gee data=armdcc;
title 'CC - GEE';
class time treat subject;
model bindif = time treat*time / noint dist=binomial;
repeated subject=subject / withinsubject=time type=exch modelse;
run;
```

Alternatively, for the linearization-based version of GEE, with empirically corrected standard errors, we can use Program 7.6.

**PROGRAM 7.6 Complete case analysis (binary outcome, GEE, linearized version, PROC GLIMMIX)**

```
proc glimmix data=armdcc empirical;
  title 'CC - GEE - linearized version - empirical';
  nloptions maxiter=50 technique=newrap;
  class time treat subject;
  model bindif = time treat*time / noint solution dist=binary;
  random _residual_ / subject=subject type=cs;
run;
```

For the generalized linear mixed model (7.5.17), with numerical quadrature, the following code is useful. See Program 7.7.

**PROGRAM 7.7 Complete case analysis (binary outcome, GLMM, PROC GLIMMIX)**

```
proc glimmix data=armdcc method=gauss(q=20);
  title 'CC - mixed - quadrature';
  nloptions maxiter=50 technique=newrap;
  class time treat subject;
  model bindif = time treat*time / noint solution dist=binary;
  random intercept / subject=subject type=un g gcorr;
run;
```

With NL MIXED, we could use Program 7.8.

**PROGRAM 7.8 Complete case analysis (binary outcome, GLMM, PROC NL MIXED)**

```
data help;
  set armdcc;
  time1=0; if time=1 then time1=1;
  time2=0; if time=2 then time2=1;
  time3=0; if time=3 then time3=1;
  time4=0; if time=4 then time4=1;
run;

proc nlmixed data=help qpoints=20 maxiter=100 technique=newrap;
  title 'CC - mixed - numerical integration';
  eta = beta11*time1+beta12*time2+beta13*time3+beta14*time4
    +b
    +(beta21*time1+beta22*time2+beta23*time3+beta24*time4)
    *(2-treat);
  p = exp(eta)/(1+exp(eta));
  model bindif ~ binary(p);
  random b ~ normal(0,tau*tau) subject=subject;
  estimate 'tau^2' tau*tau;
run;
```

Note that the DATA step in Program 7.8 merely creates dummy variables for each of the four measurement times. The ESTIMATE statement allows for the easy estimation of the random-effects variance and its standard error, because the model parameter  $\tau$  is the corresponding standard deviation.

None of the above programs is specific to CC. Only the data preprocessing using the %cc(...) macro defines it as CC.

## 7.4.2 Simple Imputation Methods

An alternative way to obtain a data set on which complete data methods can be used is filling in the missing values, instead of deleting subjects with incomplete sequences. The principle of imputation is particularly easy. The observed values are used to impute values for the missing observations. There are several ways to use the observed information. First, we can use information on the same subject (e.g., last observation carried forward). Second, information can be borrowed from other subjects (e.g., mean imputation). Finally, both within and between subject information can be used (e.g., conditional mean imputation, hot deck imputation). Standard references are Little and Rubin (2014) and Rubin (1987). Imputation strategies have historically been very popular in sample survey methods.

However, great care has to be taken with imputation strategies. Dempster and Rubin (1983) write

The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.

For example, Little and Rubin (2014) show that the method could work for a linear model with one fixed effect and one error term, but that it generally does not for hierarchical models, split-plot designs, repeated measures (with a complicated error structure), random-effects, and mixed-effects models. At the very least, different imputations for different effects would be necessary.

The user of imputation strategies faces several dangers. First, the imputation model could be wrong, and, hence, the point estimates would be biased. Second, even for a correct imputation model, the uncertainty resulting from incompleteness is masked. Indeed, even when we are reasonably sure about the mean value that the unknown observation would have, the actual stochastic realization, depending on both the mean structure as well as on the error distribution, is still unknown.

### Last Observation Carried Forward

In this case, whenever a value is missing, the last observed value is substituted. It is typically applied to settings where incompleteness is due to attrition.

Very strong and often unrealistic assumptions have to be made to ensure validity of this method. First, either when we consider a longitudinal analysis or when the scientific question is in terms of the last planned occasion, we have to believe that a subjects' measurement stays at the same level from the moment of dropout onwards (or during the period they are unobserved in the case of intermittent missingness). In a clinical trial setting, we might believe that the response profile *changes* as soon as a patient goes off treatment and even that it would flatten. However, the constant profile assumption is even stronger. Second, this method shares with other single imputation methods that it overestimates the precision by treating imputed and actually observed values on equal footing.

The situation, in which the scientific question is in terms of the last observed measurement, is often considered to be the real motivation for LOCF. However in some cases, the question, defined as such, has a very unrealistic and ad hoc flavor. Clearly, measurements at (self-selected) dropout times are lumped together with measurements made at the (investigator defined) end of the study.

### Last Observation Carried Forward and SAS

Similar steps as needed for a complete case analysis need to be performed when LOCF is the goal. For a vertically organized data set, the following macro, also written by Caroline Beunckens, can be used, in the continuous case. See Program 7.9.

#### **PROGRAM 7.9 Preparing for LOCF analysis**

```
%locf(data=armd155,id=subject,time=time,response=diff,out=armdlocf2);
```

or in the dichotomous case:

```
%locf(data=armd111,id=subject,time=time,response=bindif,out=armdlocf);
```

The arguments are exactly the same and have the same meaning as in the `%cc(...)` macro of the previous section. Note that there is now a *new* response variable created, named ‘*locf*’, which should be used in the corresponding analysis programs. Thus, all SAS procedure MIXED, GENMOD, GEE, GLIMMIX, and NLMIXED code of the previous section remains valid, upon replacing the response variables ‘*diff*’ and ‘*bindiff*’ by ‘*locf*’ and, of course, by appropriately changing the names of the data sets.

---

## 7.5 Ignorable Likelihood (Direct Likelihood)

As discussed in Section 7.3, likelihood based inference is valid whenever the mechanism is MAR and provided the technical condition holds that the parameters describing the nonresponse mechanism are distinct from the measurement model parameters (Little and Rubin, 2014). In other words, the missing data process should be ignorable in the likelihood inference sense, since then the log-likelihood partitions into two functionally independent component. As a consequence, a software module for likelihood estimation can be used, provided it can handle incompletely observed subjects. In other words, it should be able to handle subjects (or: blocks) of varying lengths, which virtually all longitudinal procedures do. The ensuing parameter estimates, standard errors, likelihood ratio values, etc. are valid.

In conclusion, a likelihood-based ignorable analysis (referred to for short as ignorable likelihood or direct likelihood) is preferable since it uses all available information, without the need to delete or to impute measurements or entire subjects. It is theoretically justified whenever the missing data mechanism is MAR. There is no statistical information distortion, given that observations are neither removed (such as in complete case analysis) nor added (such as in single imputation). There is no additional programming involved to implement an ignorable analysis in the MIXED, GLIMMIX, or NLMIXED procedures, provided the order of the measurements is correctly specified. This can be done either by supplying records with missing data in the input data set or by properly indicating the order of the measurement in the REPEATED and/or RANDOM statements.

### 7.5.1 Normally Distributed Outcomes

#### **EXAMPLE: Age-related macular degeneration trial**

We consider first a simple multivariate normal model, with unconstrained time trend under placebo, an occasion-specific treatment effect, and a  $4 \times 4$  unstructured

variance-covariance matrix. Thus,

$$Y_{ij} = \beta_{j1} + \beta_{j2}T_i + \varepsilon_{ij}, \quad (7.5.16)$$

where  $T_i = 0$  for placebo and  $T_1 = 1$  for interferon- $\alpha$ . The direct-likelihood analysis is contrasted with CC and LOCF, and parameter estimates (standard errors) for the eight mean model parameters are presented in Table 7.4.

**TABLE 7.4** The Age Related Macular Degeneration Trial. Parameter estimates (standard errors) for the linear mixed models, fitted to the continuous outcome ‘difference of the number of letters read versus baseline’. CC, LOCF, and direct likelihood.  $p$  values are presented for treatment effect at each of the four times separately, as well as for all four times jointly.

Effect	Parameter	CC	LOCF	direct lik.
Parameter estimates (standard errors)				
Intercept 4	$\beta_{11}$	-3.24(0.77)	-3.48(0.77)	-3.48(0.77)
Intercept 12	$\beta_{21}$	-4.66(1.14)	-5.72(1.09)	-5.85(1.11)
Intercept 24	$\beta_{31}$	-8.33(1.39)	-8.34(1.30)	-9.05(1.36)
Intercept 52	$\beta_{41}$	-15.13(1.73)	-14.16(1.53)	-16.21(1.67)
Treatm. eff. 4	$\beta_{12}$	2.32(1.05)	2.20(1.08)	2.20(1.08)
Treatm. eff. 12	$\beta_{22}$	2.35(1.55)	3.38(1.53)	3.51(1.55)
Treatm. eff. 24	$\beta_{32}$	2.73(1.88)	2.41(1.83)	3.03(1.89)
Treatm. eff. 52	$\beta_{42}$	4.17(2.35)	3.43(2.15)	4.86(2.31)
$p$ -values				
Treatm. eff. 4	$\beta_{12}$	0.0282	0.0432	0.0435
Treatm. eff. 12	$\beta_{22}$	0.1312	0.0287	0.0246
Treatm. eff. 24	$\beta_{32}$	0.1491	0.1891	0.1096
Treatm. eff. 52	$\beta_{42}$	0.0772	0.1119	0.0366
Treatm. eff. (overall)		0.1914	0.1699	0.1234

While there is no overall treatment effect, and the  $p$ -values between the three methods do not vary too much, the picture is different for the occasion-specific treatment effects. At week 4, all three  $p$ -values indicate significance. While this is the only significant effect when only the completers are analyzed, there is one more significant effect with LOCF (week 12) and two more when direct likelihood is employed (weeks 12 and 52). Once more, CC and LOCF miss important treatment differences, the most important one being the one at week 52, the end of the study.

### 7.5.2 Non-Gaussian Outcomes

#### EXAMPLE: Age-related macular degeneration trial

Let us now turn to a random-intercept logistic model, similar in spirit to (7.7.28):

$$\text{logit}[P(Y_{ij} = 1|T_i, t_j, b_i)] = \beta_{j1} + b_i + \beta_{j2}T_i, \quad (7.5.17)$$

with notation as before and  $b_i \sim N(0, \tau^2)$ . Both PQL and numerical integration are used for model fitting. The results for this model are given in Table 7.5.

We observe the usual downward bias in the PQL versus numerical integration analysis, as well as the usual relationship between the marginal parameters of Table 7.6 and their random-effects counterparts. Note also that the random-intercepts variance is largest under LOCF, underscoring again that this method artificially increases the association between measurements on the same subject. In this case, in contrast to marginal models, LOCF and, in fact, also CC considerably overestimate the treatment effect at certain times, by varying degrees ranging from trivial to important, in particular at 4 and 24 weeks.

**TABLE 7.5** The Age-related Macular Degeneration Trial. Parameter estimates (standard errors) for the random-intercept models: PQL and numerical-integration based fits on the CC and LOCF population, and on the observed data (direct-likelihood).

Effect	Parameter	CC	LOCF	direct lik.
PQL				
Int.4	$\beta_{11}$	-1.19(0.31)	-1.05(0.28)	-1.00(0.26)
Int.12	$\beta_{21}$	-1.05(0.31)	-1.18(0.28)	-1.19(0.28)
Int.24	$\beta_{31}$	-1.35(0.32)	-1.30(0.28)	-1.26(0.29)
Int.52	$\beta_{41}$	-1.97(0.36)	-1.89(0.31)	-2.02(0.35)
Trt.4	$\beta_{12}$	0.45(0.42)	0.24(0.39)	0.22(0.37)
Trt.12	$\beta_{22}$	0.58(0.41)	0.68(0.38)	0.71(0.37)
Trt.24	$\beta_{32}$	0.55(0.42)	0.50(0.39)	0.49(0.39)
Trt.52	$\beta_{42}$	0.44(0.47)	0.39(0.42)	0.46(0.46)
R.I. s.d.	$\tau$	1.42(0.14)	1.53(0.13)	1.40(0.13)
R.I. var.	$\tau^2$	2.03(0.39)	2.34(0.39)	1.95(0.35)
Numerical integration				
Int.4	$\beta_{11}$	-1.73(0.42)	-1.63(0.39)	-1.50(0.36)
Int.12	$\beta_{21}$	-1.53(0.41)	-1.80(0.39)	-1.73(0.37)
Int.24	$\beta_{31}$	-1.93(0.43)	-1.96(0.40)	-1.83(0.39)
Int.52	$\beta_{41}$	-2.74(0.48)	-2.76(0.44)	-2.85(0.47)
Trt.4	$\beta_{12}$	0.64(0.54)	0.38(0.52)	0.34(0.48)
Trt.12	$\beta_{22}$	0.81(0.53)	0.98(0.52)	1.00(0.49)
Trt.24	$\beta_{32}$	0.77(0.55)	0.74(0.52)	0.69(0.50)
Trt.52	$\beta_{42}$	0.60(0.59)	0.57(0.56)	0.64(0.58)
R.I. s.d.	$\tau$	2.19(0.27)	2.47(0.27)	2.20(0.25)
R.I. var.	$\tau^2$	4.80(1.17)	6.08(1.32)	4.83(1.11)

**TABLE 7.6** The Age-related Macular Degeneration Trial. Parameter estimates (model-based standard errors; empirically corrected standard errors) for the marginal models: standard and linearization-based GEE on the CC and LOCF population, and on the observed data. In the latter case, also WGEE is used. All analyses based on PROC GENMOD.

Effect	Par.	CC	LOCF	Observed data	
				Unweighted	WGEE
Standard GEE					
Int.4	$\beta_{11}$	-1.01(0.24;0.24)	-0.87(0.20;0.21)	-0.87(0.21;0.21)	-0.98(0.10;0.44)
Int.12	$\beta_{21}$	-0.89(0.24;0.24)	-0.97(0.21;0.21)	-1.01(0.21;0.21)	-1.78(0.15;0.38)
Int.24	$\beta_{31}$	-1.13(0.25;0.25)	-1.05(0.21;0.21)	-1.07(0.22;0.22)	-1.11(0.15;0.33)
Int.52	$\beta_{41}$	-1.64(0.29;0.29)	-1.51(0.24;0.24)	-1.71(0.29;0.29)	-1.72(0.25;0.39)
Tr.4	$\beta_{12}$	0.40(0.32;0.32)	0.22(0.28;0.28)	0.22(0.28;0.28)	0.80(0.15;0.67)
Tr.12	$\beta_{22}$	0.49(0.31;0.31)	0.55(0.28;0.28)	0.61(0.29;0.29)	1.87(0.19;0.61)
Tr.24	$\beta_{32}$	0.48(0.33;0.33)	0.42(0.29;0.29)	0.44(0.30;0.30)	0.73(0.20;0.52)
Tr.52	$\beta_{42}$	0.40(0.38;0.38)	0.34(0.32;0.32)	0.44(0.37;0.37)	0.74(0.31;0.52)
Corr.	$\rho$	0.39	0.44	0.39	0.33
Linearization-based GEE					
Int.4	$\beta_{11}$	-1.01(0.24;0.24)	-0.87(0.21;0.21)	-0.87(0.21;0.21)	-0.98(0.18;0.44)
Int.12	$\beta_{21}$	-0.89(0.24;0.24)	-0.97(0.21;0.21)	-1.01(0.22;0.21)	-1.78(0.26;0.42)
Int.24	$\beta_{31}$	-1.13(0.25;0.25)	-1.05(0.21;0.21)	-1.07(0.23;0.22)	-1.19(0.25;0.38)
Int.52	$\beta_{41}$	-1.64(0.29;0.29)	-1.51(0.24;0.24)	-1.71(0.29;0.29)	-1.81(0.39;0.48)
Tr.4	$\beta_{12}$	0.40(0.32;0.32)	0.22(0.28;0.28)	0.22(0.29;0.29)	0.80(0.26;0.67)
Tr.12	$\beta_{22}$	0.49(0.31;0.31)	0.55(0.28;0.28)	0.61(0.28;0.29)	1.85(0.32;0.64)
Tr.24	$\beta_{32}$	0.48(0.33;0.33)	0.42(0.29;0.29)	0.44(0.30;0.30)	0.98(0.33;0.60)
Tr.52	$\beta_{42}$	0.40(0.38;0.38)	0.34(0.32;0.32)	0.44(0.37;0.37)	0.97(0.49;0.65)
	$\sigma^2$	0.62	0.57	0.62	1.29
	$\tau^2$	0.39	0.44	0.39	1.85
Corr.	$\rho$	0.39	0.44	0.39	0.59

### 7.5.3 Direct Likelihood and SAS

In contrast to CC and LOCF, no extra data processing is necessary when a direct likelihood analysis is envisaged, provided the software tool used for analysis can handle measurement sequences of unequal length. This is the case for virtually all longitudinal data analysis tools, including the SAS procedures MIXED, NLMIXED, and GLIMMIX.

One note of caution is relevant, however. When residual correlation structures are used for which the order of the measurements within a sequence is important, such as unstructured and AR(1), but not simple or compound symmetry, and intermittent missingness occurs, care has to be taken to ensure that the *design* order within the sequence, and not the *apparent* order, is passed on. In the SAS procedure MIXED, a statement such as

```
repeated / subject=subject type=un;
```

is fine when every subject has, say, four designed measurements. However, when for a particular subject, the second measurement is missing, there is a risk that the remaining measurements are considered the first, second, and third, rather than the first, third, and fourth. Thus, it is sensible to replace the above statement by:

```
repeated time / subject=subject type=un;
```

For the GENMOD and GEE procedures, the option `withinsubject=time` of the REPEATED statement can be used. Note that this produces GEE and not direct likelihood. For the GLIMMIX procedure, there is no such feature. Evidently, we can also avoid the problem by properly sorting the measurements within a subject and at the same time ensuring that for missing values a record is included with, of course, a missing value instead of the actual measurement.

In all cases, especially when GLIMMIX is used, the proper order is passed on when a record is included, even for the missing measurements.

When the NLMIXED procedure is used, only random effects can be included. In such a case, all relevant information is contained in the actual effects that define the random effects structure. For example, the order is immaterial for a random intercepts model, and, for a random slope in time, all information needed about time is passed on, for example, by the RANDOM statement:

```
RANDOM intercept time / subject=subject type=un;
```

Thus, in conclusion, all code for likelihood-based analyses, listed in Section 7.4.1 can be used, provided the original data sets (`armd155.sas7bdat` and `armd111.sas7bdat`) are passed on, and not the derived ones.

We conclude that, with only a minimal amount of care, a direct likelihood analysis is no more complex than the corresponding analysis on a set of data that is free of missingness.

---

## 7.6 Direct Bayesian Analysis (Ignorable Bayesian Analysis)

As stated earlier, not only likelihood but also Bayesian analyses are ignorable under MAR and appropriate regularity conditions. This means that, just like with ignorable likelihood, an ignorable Bayesian analysis is as easy to carry out with

complete as well as incomplete data. To illustrate this, consider the following simple linear mixed model for the **diff** outcome:

$$Y_{ij} \sim N\left((\beta_1 + b_{1i}) + (\beta_2 + b_{2i})t_j + (\beta_3 + b_{3i})T_i + (\beta_4 + b_{4i})T_i t_j, \sigma^2\right), \quad (7.6.18)$$

$$\begin{pmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \\ b_{4i} \end{pmatrix} \sim N(\mathbf{0}, G). \quad (7.6.19)$$

To allow comparison between ignorable likelihood and ignorable Bayesian analysis, we first provide the ignorable likelihood code.

### PROGRAM 7.10 Direct likelihood

```
proc mixed data=m.armd13k method=ml nobound covtest;
title 'direct likelihood';
class subject;
model diff = time treat treat*time / solution ddfm=kr;
random intercept time treat treat*time / subject=subject type=vc;
run;
```

Relevant output is:

---

Selected direct likelihood output		Solution for Fixed Effects				
Effect	Estimate	Standard Error	DF	t Value	Pr >  t	
Intercept	4.5954	2.0798	316	2.21	0.0279	
time	-2.5089	1.0527	286	-2.38	0.0178	
treat	-1.5121	1.3584	383	-1.11	0.2664	
time*treat	-0.7952	0.6731	350	-1.18	0.2383	
Covariance Parameter Estimates						
Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z	
Intercept	subject	12.3563	12.9356	0.96	0.3395	
time	subject	14.3486	3.3803	4.24	<.0001	
treat	subject	5.1696	4.8308	1.07	0.2846	
time*treat	subject	-0.2543	1.1974	-0.21	0.8318	
Residual		50.7478	3.3151	15.31	<.0001	

---

Note that the variance component associated with the time by treatment interaction is negative, though not significantly different from zero. In other words, we have a model that allows a marginal but no hierarchical interpretation. This will be different in the upcoming Bayesian analysis, where a hierarchical interpretation is inherent in the model.

For Bayesian analysis purposes, we rewrite (7.6.18)–(7.6.19) as:

$$Y_{ij} \sim N\left(\beta_{1i} + \beta_{2i}t_j + \beta_{3i}T_i + \beta_{4i}T_i t_j, \sigma^2\right), \quad (7.6.20)$$

$$\boldsymbol{\beta}_i = \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \\ \beta_{3i} \\ \beta_{4i} \end{pmatrix} \sim N(\boldsymbol{\beta}, G), \quad (7.6.21)$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \Sigma_0), \quad (7.6.22)$$

$$G \sim \text{IWishart}(\rho, S), \quad (7.6.23)$$

$$\sigma^2 \sim \text{IGamma}(\alpha, \gamma). \quad (7.6.24)$$

The corresponding program, incorporating choices for the hyperprior parameters, is Program 7.11.

### PROGRAM 7.11 Direct Bayesian analysis

```
proc mcmc data=m.armd13k nmc=10000 outpost=m.armd131 seed=23 init=random;
title 'direct Bayes';
array theta[4] beta1 beta2 beta3 beta4;
array theta_c[4];
array dmat[4,4];
array beta0[4] (0 0 0 0);
array Sig0[4,4] (1000 0 0 0 0 1000 0 0 0 0 1000 0 0 0 0 1000);
array S[4,4] (100 0 0 0 0 100 0 0 0 0 100 0 0 0 0 100);
parms theta_c dmat {10 0 0 0 0 10 0 0 0 0 10 0 0 0 0 10} var_y;
prior theta_c ~ mvn(beta0,Sig0);
prior dmat ~ iwish(4,S);
prior var_y ~ igamma(0.01,scale=0.01);
random theta~mvn(theta_c,dmat) subject=subject;
mu=beta1+beta2*treat+beta3*time+beta4*treat*time;
model diff~normal(mu,var=var_y);
run;
```

The corresponding output is below.

---

Selected direct Bayesian output		Posterior Summaries and Intervals			
Parameter	N	Mean	Standard Deviation	95% HPD Interval	
theta_c1	10000	4.3933	2.9525	-1.8276	8.5445
theta_c2	10000	-1.4268	1.9469	-4.4090	2.5441
theta_c3	10000	-2.2187	1.2372	-4.5489	0.0692
theta_c4	10000	-0.9799	0.8187	-2.5878	0.4703
dmat1	10000	43.8501	25.5374	8.7026	97.8559
dmat2	10000	-14.0943	16.7127	-53.0171	4.6373
dmat3	10000	-13.0795	17.2618	-52.0841	11.3516
dmat4	10000	5.1822	9.6912	-8.5062	28.9990
dmat5	10000	-14.0943	16.7127	-53.0171	4.6373
dmat6	10000	21.8172	12.0049	5.2413	48.3360
dmat7	10000	0.7602	11.0049	-15.7870	28.3676
dmat8	10000	-4.1026	5.9859	-18.0130	5.5606
dmat9	10000	-13.0795	17.2618	-52.0841	11.3516
dmat10	10000	0.7602	11.0049	-15.7870	28.3676
dmat11	10000	36.4025	19.3862	9.0475	76.2131
dmat12	10000	-14.9697	11.7158	-39.6488	-0.1558
dmat13	10000	5.1822	9.6912	-8.5062	28.9990
dmat14	10000	-4.1026	5.9859	-18.0130	5.5606
dmat15	10000	-14.9697	11.7158	-39.6488	-0.1558
dmat16	10000	12.6873	6.8567	3.9227	26.8400
var_y	10000	46.2681	3.0475	40.4443	52.2954

---

We observe that the results are similar to those of the direct likelihood analysis, but that there are differences as well. This is due to the effect of the prior distributions. Another source of difference is the fact that the likelihood-based model does not impose bounds on the components of  $G$ , whereas the Bayesian model is intrinsically hierarchical.

## 7.7 Weighted Generalized Estimating Equations

---

### 7.7.1 Concept

Generalized estimating equations (GEE), as discussed in Section 7.3.4, are appealing to model repeated measures when the research questions are formulated in terms of the marginal mean function, especially but not only when outcomes are of a non-Gaussian type.

However, as Liang and Zeger (1986) pointed out, incomplete-data based inferences with GEE are valid only under the strong assumption that the data are missing completely at random (MCAR). To allow the data to be missing at random (MAR), Robins, Rotnitzky, and Zhao (1995) proposed weighted estimating equations (WGEE). In Section 7.8, we will also discuss the combination of GEE with multiple imputation.

The idea is to weight each subject's contribution to the GEE by the inverse probability, either of being fully observed, or of being observed up to a certain time. In line with Molenberghs et al. (2011), let  $\pi_i$  be the probability for subject  $i$  to be completely observed, and  $\pi'_i$  the probability for subject  $i$  to drop out on occasion  $d_i$ . These can be written as

$$\pi_i = \prod_{\ell=2}^{n_i} (1 - p_{i\ell}), \quad \pi'_i = \left[ \prod_{\ell=2}^{d_i-1} (1 - p_{i\ell}) \right] \cdot p_{id_i}, \quad (7.7.25)$$

where  $p_{i\ell} = P(D_i = \ell | D_i \geq \ell, Y_{i\bar{\ell}}, X_{i\bar{\ell}})$  are the component probabilities of dropping out at occasion  $\ell$ , given that the subject is still in the study, the covariate history  $X_{i\bar{\ell}}$ , and the outcome history  $Y_{i\bar{\ell}}$ . In such a case, we can opt either for WGEE based on the completers only:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^N \frac{\tilde{R}_i}{\pi_i} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} V_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \quad (7.7.26)$$

with  $\tilde{R}_i = 1$  if a subject is fully observed and 0 otherwise, or, upon using (7.7.25), for WGEE using all subjects:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^N \frac{1}{\pi'_i} \frac{\partial \boldsymbol{\mu}_i^o}{\partial \boldsymbol{\beta}'} (V_i^o)^{-1} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) = \mathbf{0}. \quad (7.7.27)$$

Here the superscript 'o' indicates the portion corresponding to the observed data in the corresponding matrix or vector. Of course, with (7.7.26), the incomplete subjects also contribute through the model for the dropout probabilities  $\pi_i$ .

**EXAMPLE: Age-related macular degeneration trial**

Consider the binary outcome that indicates whether the number of letters correctly read at a follow-up occasion is higher or lower than the corresponding number of letters at baseline. A population averaged (or marginal model) is used. We compare analyses performed on the completers only (CC), on the LOCF imputed data, as well as on the observed data. In all cases, standard GEE and the linearization-based version are considered. For the observed, partially incomplete data, GEE is supplemented with WGEE.

The GEE analyses are reported in Table 7.6. In all cases, we use the logit link, and the model takes the form:

$$\text{logit}[P(Y_{ij} = 1|T_i, t_j)] = \beta_{j1} + \beta_{j2}T_i, \quad (7.7.28)$$

similar in spirit to (7.5.16). A working exchangeable correlation matrix is considered. For the WGEE analysis, the following weight model is assumed:

$$\begin{aligned} & \text{logit}[P(D_i = j|D_i \geq j)] \\ &= \psi_0 + \psi_1 y_{i,j-1} + \psi_2 T_i + \psi_{31} L_{1i} + \psi_{32} L_{2i} + \psi_{34} L_{3i} \\ &+ \psi_{41} I(t_j = 2) + \psi_{42} I(t_j = 3), \end{aligned} \quad (7.7.29)$$

with  $y_{i,j-1}$  the binary outcome at the previous time  $t_{i,j-1} = t_{j-1}$ ,  $L_{ki} = 1$  if the patient's eye lesion is of level  $k = 1, \dots, 4$  (since one dummy variable is redundant, only three are used), and  $I(\cdot)$  is the indicator function. Parameter estimates and standard errors for the dropout model are given in Table 7.7. Intermittent missingness will be ignored at this time. We return to this point in Section 7.8. Covariates of importance are treatment assignment, the level of lesions at baseline (a four-point categorical variable, for which three indicator variables are needed), and time at which dropout occurs. For the latter covariates, there are three levels, since dropout can occur at times 2, 3, or 4. Hence, two indicator variables are included. Finally, the previous outcome does not have a significant impact, but will be kept in the model nevertheless.

**TABLE 7.7** The Age-related Macular Degeneration Trial. Parameter estimates (standard errors) for a logistic regression model to describe dropout.

Effect	Parameter	Estimate (s.e.)	
		GENMOD	GEE
Intercept	$\psi_0$	0.14 (0.49)	0.17 (0.56)
Previous outcome	$\psi_1$	0.04 (0.38)	-0.05 (0.38)
Treatment	$\psi_2$	-0.86 (0.37)	-0.87 (0.37)
Lesion level 1	$\psi_{31}$	-1.85 (0.49)	-1.82 (0.49)
Lesion level 2	$\psi_{32}$	-1.91 (0.52)	-1.88 (0.52)
Lesion level 3	$\psi_{33}$	-2.80 (0.72)	-2.79 (0.72)
Time 2	$\psi_{41}$	-1.75 (0.49)	-1.73 (0.49)
Time 3	$\psi_{42}$	-1.38 (0.44)	-1.36 (0.44)

*Note: GENMOD is called after the %dropout macro is called. The GEE parameters result from the MISSMODEL statement within the procedure.*

From Table 7.6, it is clear that there is very little difference between the standard GEE and linearization-based GEE results. This is undoubtedly the case for CC, LOCF, and unweighted GEE on the observed data. For these three cases, also, the model-based and empirically corrected standard errors agree extremely well, owing to the unstructured nature of the full time by treatment mean structure. However, we do observe differences in the WGEE analyses. Not only do the parameter estimates

differ a little between the two GEE versions, but there is also a dramatic difference between the model-based and empirically corrected standard errors. This is entirely due to the weighting scheme. The weights were not calibrated to add up to the total sample size, which is reflected in the model-based standard errors. In the linearization-based case, part of the effect is captured as overdispersion. This can be seen from adding the parameters  $\sigma^2$  and  $\tau^2$ . In all other analyses, the sum is close to one, as it should be when there is no residual overdispersion, but, in the last column, these add up to 3.14. Nevertheless, the two sets of empirically corrected standard errors agree very closely, which is reassuring.

In spite of there being no strong evidence for MAR, the results between GEE and WGEE differ in a nontrivial way. It is noteworthy that at 12 weeks, a treatment effect is observed with WGEE, which is undetected when using the other marginal analyses. This finding is confirmed to some extent by the subject-specific random-intercept model, presented in the next section, when the data are used as observed.

When comparing parameter estimates across CC, LOCF, and observed data analyses, it is clear that LOCF has the effect of artificially increasing the correlation between measurements. The effect is mild in this case. The parameter estimates of the observed-data GEE are close to the LOCF results for earlier time points and close to CC for later time points. This is to be expected, as at the start of the study the LOCF and observed populations are very similar, with the same holding between CC and observed populations near the end of the study. Note also that the treatment effect under LOCF, especially at 12 weeks and after 1 year, is biased downward in comparison to the GEE analyses.

### **7.7.2 WGEE and SAS, Using PROC GENMOD**

We will first discuss the steps to be taken when using the older SAS procedure GENMOD. Afterwards, we will switch to the more recent and easier to use GEE procedure.

A GENMOD program for the standard GEE analysis is Program 7.12.

#### **PROGRAM 7.12 Standard GEE**

```
proc genmod data=armdwgee;
class time treat subject;
model bindif = time treat*time / noint dist=binomial;
repeated subject=subject / withinsubject=time type=exch modelse;
run;
```

Likewise, the linearization-based version can be used without any problem, using Program 7.13:

#### **PROGRAM 7.13 Linearization-based GEE**

```
proc glimmix data=armdwgee empirical;
nloptions maxiter=50 technique=newrap;
class time treat subject;
model bindif = time treat*time / noint solution dist=binary;
random _residual_ / subject=subject type=cs;
run;
```

Note that PROC GENMOD produces empirical as well as a model-based standard errors simultaneously, because of the `modelse` option. In the GLIMMIX code, we merely obtain the empirically corrected standard errors, because of the `empirical` option. Upon omitting this option, the model-based standard errors are obtained.

We now sketch the steps to be taken when conducting a weighted GEE analysis. To compute the weights, we first have to fit the dropout model using, for example, logistic regression. The outcome `dropout` is binary and indicates whether dropout occurs at a given time from the start of the measurement sequence until the time of dropout or the end of the sequence. Covariates in the model are the outcomes at previous occasions (`prev`), supplemented with genuine covariate information. The `%dropout` macro, constructed by Caroline Beunckens, is used to construct the variables `dropout` and `prev`.

Likewise, once a logistic regression has been fitted, these need to be translated into weights. These weights are defined at the individual measurement level and are equal to the product of the probabilities of not dropping out up to the measurement occasion. The last factor is either the probability of dropping out at that time or continuing the study. This task can be performed with the `%dropwgt` macro. The arguments are the same as in the `%dropout` macro, except that now also the predicted values from the logistic regression have to be passed on through the `pred=` argument, and the dropout indicator is passed on through the `dropout=` argument.

Using these macros, Program 7.14 can be used to prepare for a WGEE analysis.

#### **PROGRAM 7.14 Preparing for WGEE (PROC GENMOD)**

```
%dropout(data=armd111,id=subject,time=time,response=bindif,out=armdhlp);

proc genmod data=armdhlp descending;
class trt prev lesion time;
model dropout = prev trt lesion time / pred dist=binomial;
ods output obstats=pred;
run;

data pred;
set pred;
keep observation pred;
run;

data armdhlp;
merge pred armdhlp;
run;

%dropwgt(data=armdhlp,id=subject,time=time,pred=pred,
dropout=dropout,out=armdwgee);
```

To sum up, the dropout indicator and previous outcome variable are defined using the `%dropout` macro, after an ordinary logistic regression is performed. Predicted values are first saved and then merged with the original data. Finally, the predicted values are translated into proper weights using the `%dropwgt` macro. Note that this approach is restricted to subject-level weights.

After these preparatory steps, we need only include the weights through the `WEIGHT` (or, equivalently, `SCWGT`) statement within the `GENMOD` procedure. This statement identifies a variable in the input data set to be used as the exponential family dispersion parameter weight for each observation. The exponential family dispersion parameter is divided by the `WEIGHT` variable value for each observation. Whereas the inclusion of the `REPEATED` statement turns a univariate exponential family model into GEE, the addition of `WEIGHT` further switches to WGEE. In other words, we merely need to add Program 7.15.

#### **PROGRAM 7.15 Additional statement for weighted generalized estimating equations**

```
weight wi;
```

Note that the use of the WEIGHT statement can also be used in the GLIMMIX procedure, so implying that a weighted version of the linearization-based GEE method is feasible.

### 7.7.3 WGEE and SAS, Using PROC GEE

The standard GEE analysis as presented in Program 7.12 can equivalently be implemented using PROC GEE. Literally, all it takes is to replace ‘GENMOD’ by ‘GEE’. The main value of the procedure lies in the ease with which WGEE can be conducted. The preparatory steps are limited to defining the additional variables, needed in the weight model. In this case, this is the variable `prevbindif`, containing the previous value of `bindif` and indicators for the second and third time point.

#### **PROGRAM 7.16 Preparing for WGEE (PROC GEE)**

```
data help;
set armdwgee;
by subject;
prevbindif=lag(bindif);
if first.id then prevbindif=1;
time2=0;
if time=2 then time2=1;
time3=0;
if time=3 then time3=1;
run;
```

Upon completing this step, we merely need to add the MISSMODEL statement to the PROC GEE code. There is no need to specify an outcome variable, because this will always be the dropout indicator.

#### **PROGRAM 7.17 WGEE, Using PROC GEE**

```
proc gee data=help;
class time treat subject lesion;
model bindif = time treat*time / noint dist=binomial;
repeated subject=subject / withinsubject=time type=exch corrw modelse;
missmodel prevbindif treat lesion time2 time3 / type=obslevel;
run;
```

Note that we also include `type=obslevel` as an option to the MISSMODEL statement. It specifies that the weights need to be calculated at the level of the observation, rather than at the level of the subject as a whole. The latter corresponds to the `type=sublevel` option. Observation-level weights are the default. The estimates for the weight model are presented in Table 7.7 (second column). Note that they are similar but not identical to the ones obtained from the earlier analysis. The reason is that in the GENMOD-based analysis, all usable information from the non-monotone sequences is used to compute the weights, whereas in the PROC GEE analysis, non-monotone subjects are removed entirely from analysis. This also explains the small differences between the results presented in Tables 7.6 and 7.8. The Standard GEE (WGEE) analysis in the former is comparable to the subject-level analysis in the latter.

A very striking feature is the difference between observation- and subject-level weighting in Table 7.8. Some standard errors are 50–100% larger when observation-level weights are employed. In other words, such more refined weights overcome one of the main issues with WGEE, i.e., that of reduced precision. To emphasize this,

**TABLE 7.8** The Age-related Macular Degeneration Trial. Parameter estimates (empirically corrected standard errors) for WGEE using PROC GEE, with both observation-level weights (observation) and subject-level weights (subject).

Effect	Parameter	Weights	
		observation	subject
Int.4	$\beta_{11}$	-0.95 (0.20)	-0.98 (0.35)
Int.12	$\beta_{21}$	-1.03 (0.22)	-1.77 (0.30)
Int.24	$\beta_{31}$	-1.03 (0.23)	-1.11 (0.29)
Int.52	$\beta_{41}$	-1.52 (0.30)	-1.72 (0.37)
Tr.4	$\beta_{12}$	0.32 (0.28)	0.78 (0.56)
Tr.12	$\beta_{22}$	0.65 (0.29)	1.83 (0.47)
Tr.24	$\beta_{32}$	0.39 (0.30)	0.71 (0.49)
Tr.52	$\beta_{42}$	0.30 (0.39)	0.72 (0.47)
Corr.	$\rho$	0.38	0.33

the observation-level analysis produces standard errors in the order of magnitude of unweighted GEE (Table 7.8).

### 7.7.4 Double Robustness

We finish the section on GEE by referring to more recent developments by Robins and colleagues that are designed to improve the efficiency of WGEE, more generally termed inverse probability weighting (IPW). For overviews, see Carpenter, Kenward, and Vansteelandt (2006); Molenberghs and Kenward (2007); and Molenberghs et al. (2015). Essentially, standard WGEE are supplemented with a second term, which has expectation zero given the observed data, and is most often written in terms of the predictive distribution of the unobserved outcomes given the observed ones. The method is termed doubly robust because it leads to a consistent and asymptotically normal estimator when either the model for the weights or the predictive model is correctly specified, but not necessarily both. Currently, the methodology is not yet available in standard procedures, although various user-defined implementations exist. See, for example (at [www.missingdata.org.uk.](http://www.missingdata.org.uk/)), Mallinckrodt and Lipkovich (2016, Ch. 17), which presents SAS code for double robust estimation.

---

## 7.8 Multiple Imputation

---

Next to the methods already discussed, multiple imputation (MI) is an attractive tool in the modeler's kit. The method is ignorable under MAR. Extensions exist for MNAR mechanisms. These are well suited for sensitivity analyses and will be discussed in Section 7.11.

Multiple imputation (MI) was formally introduced by Rubin (1978). Rubin (1987) provides a comprehensive early treatment. Several other sources, such as Rubin and Schenker (1986); Little and Rubin (2014); Tanner and Wong (1987); Schafer (1997); van Buuren (2012); Carpenter and Kenward (2013); and O'Kelly and Ratitch (2014), offer easy-to-read descriptions of the technique.

The key MI idea is to replace each missing value with a set of  $M$  plausible values, i.e., values "drawn" from the distribution of our data, that represent the uncertainty about the right value to impute. The imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these

analyses. Drawing imputations can be done in a large variety of ways, and the most commonly used of them will be discussed in what follows.

An evident question is when to use multiple imputation. This question is relevant because, given the availability of such procedures as MIXED, GLIMMIX, NLMIXED, MCMC, and related software tools, direct likelihood and direct Bayesian analyses are within reach. Also WGEE has become relatively easy to use thanks to the GEE procedures.

That said, we broadly see at least six settings where MI can be of use, without limiting it to other uses. First, when there is a combination of missing covariates and missing outcomes, multiple imputation comes in handy to either handle the incomplete covariates, or the incomplete covariates and outcomes combined. In the former case, a standard missing-data technique can be used on the incomplete outcome data with completed covariates. In the latter one, any complete-data technique can be used. Second, when several analyses are envisaged on the same set of incomplete data, missingness can be handled using MI, after which the various analyses can be undertaken. A simple example is when the same set of incomplete data will be modeled using both GEE and GLMM. Third, when a technique requires the missing data patterns to be monotone, MI can be used, either as an alternative to the technique envisaged (e.g., no direct likelihood or WGEE but rather a complete-data technique after MI), or to monotonize the incomplete data (e.g., enabling WGEE). Fourth, MI is attractive when likelihood-based analyses turn out to be difficult to implement, especially when we would like to use jointly several outcomes (e.g., binary, continuous, and/or count outcomes). See also Mallinckrodt and Lipkovich (2016, Sec. 15.4). Fifth, certain MI extensions are applicable when MNAR-type analyses are considered, in particular in the context of sensitivity analysis (see Section 7.11). Sixth, MI is useful when an analysis is to be conducted based on a discretized version of an incompletely observed continuous outcome or set of outcomes. We can then begin by imputing the original, continuous outcome, followed by discretizing the so-obtained completed data sets.

Technically, MI involves three distinct steps:

**Imputation step.** The missing values are filled in  $M$  times to generate  $M$  complete data sets.

**Analysis step.** The  $M$  complete data sets are analyzed by using standard procedures.

**Inference step.** The results from the  $M$  analyses are combined for inference purposes.

The SAS procedure MI creates multiple imputed data sets from incomplete  $p$ -dimensional multivariate data. It uses methods that incorporate appropriate variability across the  $M$  imputations. Once the  $M$  complete data sets are analyzed by using standard procedures, PROC MIANALYZE can be used to generate valid statistical inferences about these parameters by combining results for the  $M$  complete data sets. Alternative versions exist to combine, for example,  $M$   $p$ -values into a single one. More details on SAS for MI are provided in Sections 7.8.5--7.8.6.

### 7.8.1 Theoretical Justification

Suppose we have a sample of  $N$ , i.i.d.  $n \times 1$  random vectors  $\mathbf{Y}_i$ . In a data set with the number of measurements per subject  $n_i$  variable, we can define  $n = \max_{i=1}^N n_i$ . Our interest lies in estimating some parameter vector  $\boldsymbol{\theta}$  of the distribution of  $\mathbf{Y}_i$ . Multiple imputation fills in the missing data  $\mathbf{Y}^m$  using the observed data  $\mathbf{Y}^o$ , several times, and then the completed data are used to estimate  $\boldsymbol{\theta}$ .

If we knew the distribution of  $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$ , with parameter vector  $\boldsymbol{\theta}$ , then we could impute  $\mathbf{Y}_i^m$  by drawing a value of  $\mathbf{Y}_i^m$  from the conditional distribution

$$f(\mathbf{y}_i^m | \mathbf{y}_i^o, \boldsymbol{\theta}).$$

The objective of the imputation process is to sample from this true predictive distribution. Since we do not know  $\boldsymbol{\theta}$ , we must estimate it from the data, say  $\hat{\boldsymbol{\theta}}$ , and presumably use

$$f(\mathbf{y}_i^m | \mathbf{y}_i^o, \hat{\boldsymbol{\theta}})$$

to impute the missing data. Frequentists sometimes favor incorporating uncertainty in  $\boldsymbol{\theta}$  in the multiple imputation scheme using bootstrap or other methods. However, in Bayesian terms,  $\boldsymbol{\theta}$  is a random variable, in which the posterior distribution is a function of the data, so we must account for its uncertainty. The Bayesian approach relies on integrating out  $\boldsymbol{\theta}$ , which provides a more natural and unifying framework for accounting for the uncertainty in  $\boldsymbol{\theta}$ . Thus,  $\boldsymbol{\theta}$  is a random variable with mean equal to the estimated  $\hat{\boldsymbol{\theta}}$  from the data. Given this distribution, using multiple imputation, we first draw a random  $\boldsymbol{\theta}^*$  from the distribution of  $\boldsymbol{\theta}$ , and then put this  $\boldsymbol{\theta}^*$  in to draw a random  $\mathbf{Y}_i^m$  from

$$f(\mathbf{y}_i^m | \mathbf{y}_i^o, \boldsymbol{\theta}^*).$$

The imputation algorithm is as follows:

1. Draw  $\boldsymbol{\theta}^*$  from the distribution of  $\boldsymbol{\theta}$ .
2. Draw  $\mathbf{Y}_i^{m*}$  from  $f(\mathbf{y}_i^m | \mathbf{y}_i^o, \boldsymbol{\theta}^*)$ . This can be done in a variety of ways, including multivariate normal models, log-linear models, a combination thereof, Monte Carlo Markov chain methods, so-called full conditional specification (FCS), etc. (van Buuren, 2012; Carpenter and Kenward, 2013).
3. To estimate  $\boldsymbol{\beta}$ , we then calculate the estimate of the parameter of interest, and its estimated variance, using the completed data,  $(\mathbf{Y}^o, \mathbf{Y}^{m*})$ :

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y}) = \hat{\boldsymbol{\beta}}(\mathbf{Y}^o, \mathbf{Y}^{m*}),$$

and the *within*-imputation variance is  $\mathbf{U} = \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ .

4. Repeat steps 1, 2, and 3 a number of  $M$  times  $\Rightarrow \hat{\boldsymbol{\beta}}^m$  and  $\mathbf{U}^m$ , for  $m = 1, \dots, M$ .

Steps 1 and 2 constitute the *Imputation Task*. Step 3 is the *Analysis Task*.

### 7.8.2 Pooling Information

Of course, we want to combine the  $M$  inferences into a single one (the *Inference Task*). In this section, we will discuss parameter and precision estimation.

With no missing data, suppose that inference about the parameter  $\boldsymbol{\beta}$  is made by

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \sim N(\mathbf{0}, \mathbf{U}).$$

The  $M$  within-imputation estimates for  $\boldsymbol{\beta}$  are pooled to give the multiple imputation estimate

$$\hat{\boldsymbol{\beta}}^* = \frac{\sum_{m=1}^M \hat{\boldsymbol{\beta}}^m}{M}.$$

Further, we can make normal based inferences for  $\boldsymbol{\beta}$  based upon

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^*) \sim N(\mathbf{0}, \mathbf{V}),$$

where

$$\mathbf{V} = \mathbf{W} + \left( \frac{M+1}{M} \right) \mathbf{B},$$

$$\mathbf{W} = \frac{\sum_{m=1}^M \mathbf{U}^m}{M}$$

is the average within-imputation variance, and

$$\mathbf{B} = \frac{\sum_{m=1}^M (\hat{\boldsymbol{\beta}}^m - \hat{\boldsymbol{\beta}}^*) (\hat{\boldsymbol{\beta}}^m - \hat{\boldsymbol{\beta}}^*)'}{M-1}.$$

is the *between*-imputation variance (Rubin, 1987).

### 7.8.3 Hypothesis Testing

When MI is used, the asymptotic results and, hence, the  $\chi^2$  reference distributions do not only depend on the sample size  $N$ , but also on the number of imputations  $M$ . Therefore, Li, Raghunathan, and Rubin (1991) propose the use of an  $F$  reference distribution with appropriate degrees-of-freedom. To test the hypothesis  $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ , they advocate the following method to calculate  $p$ -values:

$$p = P(F_{k,w} > F),$$

where  $k$  is the length of the parameter vector  $\boldsymbol{\theta}$ ,  $F_{k,w}$  is an  $F$  random variable with  $k$  numerator and  $w$  denominator degrees of freedom, and

$$\begin{aligned} F &= \frac{(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)' W^{-1} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)}{k(1+r)}, \\ w &= 4 + (\tau - 4) \left[ 1 + \frac{(1 - 2\tau^{-1})}{r} \right]^2, \\ r &= \frac{1}{k} \left( 1 + \frac{1}{M} \right) \text{tr}(BW^{-1}), \\ \tau &= k(M-1). \end{aligned}$$

Here,  $r$  is the average relative increase in variance due to nonresponse across the components of  $\boldsymbol{\theta}$ . The limiting behavior of this  $F$  variable is that if  $M \rightarrow \infty$ , then the reference distribution of  $F$  approaches an  $F_{k,\infty} = \chi^2/k$  distribution.

Clearly, this procedure is not only applicable when the full vector  $\boldsymbol{\theta}$ , but also when one component, a sub-vector, or a set of linear contrasts, is the subject of hypothesis testing. In case of a sub-vector, or as a special case one component, we use the corresponding sub-matrices of  $B$  and  $W$  in the formulas. For a set of linear contrasts  $L\boldsymbol{\beta}$ , we should use the appropriately transformed covariance matrices:  $\tilde{W} = LWL'$ ,  $\tilde{B} = LBL'$ , and  $\tilde{V} = LVL'$ .

### 7.8.4 Efficiency

Multiple imputation is attractive because it can be highly efficient even for small values of  $M$ . Historically, numbers as small as  $M = 5$  were often advocated (Rubin, 1987, p. 114). Of course, efficiency depends on a variety of factors, such as the amount of missingness, data type, and whether the inferential goal is estimation or

hypothesis testing. It is prudent to use somewhat higher values. The current SAS default is  $M = 25$  (as of SAS/STAT 14.1; formerly  $M = 5$  was the default). Users are encouraged to conduct simple numerical sensitivity analyses, by varying the number of imputations over a range of values, until a desired level of precision is attained. Carpenter and Kenward (2013) offer guidelines in this respect.

### 7.8.5 Imputation Mechanisms

The method of choice to create the imputed data sets depends on the missing data pattern and the type(s) of the outcome variables. Carpenter and Kenward (2013) describe the most commonly available methods for univariate and multivariate outcomes, as well as a number of methods developed for specific cases such as time-to-event data, data with nonlinear relationships, multilevel models, etc.

A data set is said to have a monotone missing data pattern if, a missing outcome  $Y_{ij}$  implies that  $Y_{ik}$ ,  $k > j$  are missing for the same individual  $i$ , perhaps after permuting the columns of the data matrix with components  $Y_{ij}$ .

The widest array of methods is available for monotone data. In the SAS procedure MI, apart from the general MCMC and FCS statements, also the MONOTONE statement can be used. Within these, several options, and hence methods, are available.

For monotone missing data patterns, a parametric *regression method* can be used that assumes multivariate normality, logistic regression, or a combination thereof. When opting for this approach, a regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the resulting model, a new regression model is then fitted and is used to impute the missing values for each variable (Rubin, 1987), in line with steps 1 and 2 in Section 7.8.1. Since the data set has a monotone missing data pattern, the process can easily be repeated sequentially for variables with missing values. To this end, the options `reg` and `logistic` are available. In addition, for categorical data, a discriminant analysis based method can be used, using the `discrim` option.

When the `logistic` option is used with the FCS or MONOTONE statements, we can make use of the `likelihood=augment` sub-option. This tool is handy when maximum likelihood estimates for logistic regression do not exist (or tend to infinity) because of so-called quasi-complete separation. The methodology was developed by White, Daniel, and Royston (2010).

Alternatively, we can rely on a nonparametric method that uses *propensity scores*, by means of the `propensity` option. The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates (Rosenbaum and Rubin, 1983). In the propensity score method, a propensity score is generated for each variable with missing values to indicate the probability of that observation being missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation (Rubin, 1987) is applied to each group. The propensity score method uses only the covariate information that is associated with whether the imputed variable values are missing. It does not use correlations among repeated measures. It is effective for inferences about the distributions of individual imputed variables, but it is not appropriate for analyses involving relationships among variables.

Finally, for monotone data, the so-called *predictive mean matching (PMM)* is also available, using the option `regpmm`. The method is similar to regression imputation, except that the value imputed is not merely taken from the predictive distribution, but rather from a pool of donors with value close to the predictive mean. In some applications, simply the closest value is selected, while in others a random selection still takes place (Heitjan and Litte, 1991; Carpenter and Kenward, 2013, p. 133).

For arbitrary missing data patterns, and, hence, in particular for monotone patterns as well, there are two main methods: *multivariate modeling* (also referred to as joint modeling; Schafer 1997, Carpenter and Kenward 2013) and full conditional specification (FCS). In SAS, the former is implemented using the MCMC statement, while the latter has become available since SAS 9.4 by way of the FCS statement.

In statistical applications, MCMC is used to generate pseudo-random draws from multidimensional and otherwise intractable probability distributions via Markov chains. A Markov chain is a sequence of random variables in which the distribution of each element depends on the value of the previous one(s). In the *MCMC method*, we construct a Markov chain long enough for the distribution of the elements to stabilize to a target distribution. This stationary distribution is the one of interest. By repeatedly simulating steps of the chain, it simulates draws from the distribution of interest.

In more detail, the MCMC method works as follows. We assume that the data are from a multivariate normal distribution,  $N(\boldsymbol{\mu}, \Sigma)$ , say. We then proceed as follows.

1. In the first step, the **initial step**, we have to choose starting values,  $\mathbf{Y}_i \sim \boldsymbol{\mu}^{(0)}$  and  $\Sigma^{(0)}$ , say. This can be done by computing a vector of means and a covariance matrix from the complete data. These are used to estimate the prior distribution. More precisely, this means that the parameters of the prior distributions for means and variances of the multivariate normal distribution are estimated, using the informative prior option.
2. The next step is called the **imputation step**: Values for missing data items are simulated by randomly selecting a value from the available distribution of values, i.e., the predictive distribution of missing values given the observed values. Technically, the predictive distribution makes use of the result:

$$\begin{aligned}\mathbf{Y}_i^{(m)} | \mathbf{y}_i^{(o)} \\ \sim N\left(\boldsymbol{\mu}_i^{(m)} + \Sigma^{(mo)} \Sigma^{(mm)-1} \left(\mathbf{y}_i^{(o)} - \boldsymbol{\mu}_i^{(o)}\right); \Sigma^{(mm)} - \Sigma^{(mo)} \Sigma^{(mm)-1} \Sigma^{(mo)}\right).\end{aligned}$$

3. In the **posterior step**, the posterior distribution of the mean and covariance parameters is updated, by updating the parameters governing their distribution (e.g., the inverted Wishart distribution for the variance-covariance matrix and the normal distribution for the means). This is then followed by sampling from the posterior distribution of mean and covariance parameters, based on the updated parameters.
4. The previous two steps, i.e., the imputation and the posterior steps, are iterated until the distribution is stationary. This means that the mean vector and covariance matrix are unchanged as we iterate.
5. To conclude, we use the imputations from the final iteration to form a data set that has no missing values.

The MCMC method is implemented in the SAS procedure MI via the MCMC statement. While in principle the method applies to arbitrary data type, its SAS implementation is restricted to multivariate normal data. If needed, the TRANSFORM statement can be used to transform clearly non-normal outcomes to (near) normality. An important feature of the MCMC statement is that it can be used to either impute all missing values or just enough to make non-monotonically missing data patterns monotone. In the latter case, the ‘impute=monotone’ option has to be used. It is important to note that this option is entirely different, syntactically and semantically, from the MONOTONE statement discussed earlier:

- The MONOTONE statement takes monotone patterns as input and returns completed data.

- The MCMC statement with `impute=monotone` option also takes data with non-monotone patterns as input and returns monotonized data.

The most recent addition to the MI procedure is FCS, through the FCS statement (van Buuren et al., 1999; van Buuren, 2007, 2012; and Carpenter and Kenward, 2013, Sec. 3.3). The monotone regression method is easy, because every missing value in a monotone sequence can be predicted by its predecessors. It is flexible in that it can handle a combination of continuous and categorical outcomes. On the other hand, many application data sets have non-monotone patterns, even though they might not be the majority of the patterns. The method can best be viewed as a modification of the monotone method. The outcomes are first ordered such that the patterns are as close to monotone as possible. As an initial step, the missing values are imputed by drawing with replacement from the observed values of each variable. Then a number of cycles are run, each of which consists of two steps:

- A regression of the observed part of the  $j$ th variable,  $\mathbf{Y}_j$ , on the remaining variables is conducted. In this regression, missing values are replaced by the current value of the imputations.
- Given the result of these regressions, and using the same algorithm as with regression imputation, new imputations for the missing values in  $\mathbf{Y}_j$  are imputed. To account for parameter uncertainty, not just outcomes but also the parameters  $\boldsymbol{\theta}$  are sampled.

After a set of burn-in sequences, the first imputation is obtained. After that, a new set of cycles is run to obtain the second imputation, etc. Apart from the regression method, the FCS statement also allows for the logistic, discriminant, and predictive mean matching methods.

#### **EXAMPLE: Age-related macular degeneration trial**

When analyzing the data using WGEE in Section 7.7, one complication that arose was that only monotone missingness is allowed. We, therefore, removed the non-monotone sequences. To overcome this, MI is an appealing alternative. We can either monotonize the data and still apply WGEE, or impute the incomplete data altogether, followed by standard GEE. We will refer to the latter methods as MI-GEE.

An appealing feature of MI is that imputation can be based on the continuous outcome (visual acuity in this case) *before* dichotomizing the outcome. See also Mallinckrodt and Lipkovich (2016, Sec. 15.4). In other words, more information can be used during imputation than when analyzing the data. Here, this takes the form of outcomes prior to dichotomization. Additionally, auxiliary covariates can be used in the imputation process as well. We take both of these measures in this analysis. Ten multiply-imputed data sets were created. The imputation model also included, apart from the four continuous outcome variables, the four-point categorical variable ‘lesions.’ For simplicity, the latter was treated as continuous. Separate imputations were conducted for each of the two treatment groups. These choices imply that the imputed values depend on lesions and treatment assignment, and, hence, analysis models that include one or both of these effects are *proper* in the sense of Rubin (1987). This means, broadly speaking, that the model used for imputation should include all relationships that will be considered later in the analysis and inference steps. The added advantage of including ‘lesions’ in the imputation model is that even individuals for which none of the four follow-up measurements are available, are still imputed and hence retained for analysis. Using the SAS procedure MI, the MCMC method was used, with EM starting values, and a single chain for all imputations.

Upon imputation, a marginal model (using GEE as in Section 7.7) was fitted, together with a generalized linear mixed model as in Section 7.5. The final results, obtained by making use of Rubin's combination rules, are reported in Table 7.9. The parameter estimates and standard errors are very similar to their counterparts in Table 7.6 and 7.5, respectively. In the GEE case, the parameter estimates are similar to those in Table 7.8, obtained with observation-level weights. Also, the similarity between the direct likelihood method (bottom right column of Table 7.5) is clear, with only a minor deviation in estimate for the treatment effect after one year.

**TABLE 7.9** The Age-related Macular Degeneration Trial. Parameter estimates (standard errors) for the standard GEE and numerical-integration based random-intercept models, after generating 10 multiple imputations.

Effect	Par.	GEE	GLMM
Int.4	$\beta_{11}$	-0.84(0.20)	-1.46(0.36)
Int.12	$\beta_{21}$	-1.02(0.22)	-1.75(0.38)
Int.24	$\beta_{31}$	-1.07(0.23)	-1.83(0.38)
Int.52	$\beta_{41}$	-1.61(0.27)	-2.69(0.45)
Trt.4	$\beta_{12}$	0.21(0.28)	0.32(0.48)
Trt.12	$\beta_{22}$	0.60(0.29)	0.99(0.49)
Trt.24	$\beta_{32}$	0.43(0.30)	0.67(0.51)
Trt.52	$\beta_{42}$	0.37(0.35)	0.52(0.56)
R.I. s.d.	$\tau$		2.20(0.26)
R.I. var.	$\tau^2$		4.85(1.13)

## 7.8.6 SAS for Multiple Imputation

To conduct multiple imputation in SAS, a sequence of three procedures is used:

**Imputation Task: PROC MI:** To generate  $M$  imputed data sets.

**Analysis Task: Data analysis procedure:** Using an appropriate procedure or other analysis tool, the  $M$  imputed data sets are analyzed. For example, if a GLMM is envisaged, PROC GLIMMIX or PROC NLMIXED can be used. Routinely, parameter estimates and their estimated variance-covariance matrices are saved into data sets.

**Inference Task: PROC MIANALYZE:** The combination rules are applied to the data sets saved in the previous step and appropriate inferences are drawn.

We discuss each of these in turn.

### PROC MI

Some information on PROC MI was already given in Section 7.8.5, in relation to the methodology used for generating imputations.

PROC MI creates  $M$  imputed data sets, physically stored in a single data set with indicator `_IMPUTATION_` to separate the various imputed copies from each other. We will describe some options available in the PROC MI statement. The option `simple` displays simple descriptive statistics and pairwise correlations based on available cases in the input data set. The number of imputations is specified by `nimpute` and is by default equal to 25 (as of SAS/STAT 14.1). The option `round` controls the number of decimal places in the imputed values (by default, there is

no rounding). If more than one number is specified, we should also use a VAR statement, and the specified numbers must correspond to variables in the VAR statement. The **seed** option specifies a positive integer, which is used by PROC MI to start the pseudo-random number generator. The default is a value generated from the time of day from the computer's clock.

The imputation step is carried out separately for each level of the BY variables.

As stated in Section 7.8.5, there are three imputation statements. For monotone missingness only, we use the MONOTONE statement, with options: **reg** for the standard regression method, **logistic** for the logistic regression method, **discrim** for the discriminant analysis method, **regpmm** for the predictive mean matching method, and **propensity** for the propensity score method. We can specify more than one method in the MONOTONE statement, and for each imputed variable, the covariates can be specified separately.

For general patterns of missingness, we can use the MCMC statement, which is also the default. Recall that the method uses a multivariate normal model, or the MCMC method. We can give the initial mean and covariance estimates to begin the MCMC process by **initial**. Tools are available to monitor convergence of the MCMM sequence. With **initial=EM** (default), PROC MI uses the means and standard deviations from available cases as the initial estimates for the EM algorithm. The resulting estimates are used to begin the MCMC process. We can also specify **initial=input SAS-data-set** to use a SAS data set with the initial estimates of the mean and covariance matrix for each imputation. Further, **niter** specifies the number of iterations between imputations in a single chain (the default is equal to 30).

As already mentioned in Section 7.8.5, for full conditional specification, the FCS statement is available, with the same modeling options as for the MONOTONE statement (**reg**, **logistic**, **discrim**, and **regpmm**), except **propensity**. Using **nbiter**, the number of burn-in iterations can be specified.

The CLASS statement is intended to specify categorical variables. Such classification variables are used as either covariates for imputed variables or as imputed variables for data sets with monotone missingness patterns. When a CLASS statement is included, either MONOTONE or FCS must be used.

Of note is the EM statement. It calculates expectation-maximization (EM) algorithm based parameter estimates for a multivariate normal sample (Dempster, Laird, and Rubin, 1977). When the number of iterations is set equal to zero and the EM statement is invoked, PROC MI in fact calculates EM-based rather than MI-based estimates. The EM estimates are also useful as initial values for the various MI techniques.

When a variable is assumed to be normally distributed but its actual distribution is deviating from it, the TRANSFORM statement can be used to transform the variable, using one of the prescribed transformations, such as Box-Cox, logarithmic, logistic, etc.

When using MI for longitudinal or otherwise hierarchical data, some data analysis is necessary before and after invoking PROC MI. In most hierarchical data sets, there is a single data set line reserved for each measurements. This implies that a subject (block) runs across several lines. We term this the vertical (counting process) layout. However, PROC MI assumes that each line is an independent block, the horizontal (multivariate) layout. Therefore, a hierarchical data set has to be transformed from a vertical to a horizontal format prior to calling PROC MI. Afterwards, the output data set needs to be transformed again to the vertical format, to allow calling one of the hierarchical procedures, such as GENMOD, GEE, MIXED, GLIMMIX, NLMIXED, etc.

Discussion of one further statement available in the MI procedure, the MNAR statement, is deferred to Section 7.11.

### **Data Analysis Procedure**

Next, the imputed data sets are analyzed using a standard procedure. It is important to ensure that the `BY _imputation_` syntax is used to force an analysis for each of the imputed sets of data separately. Appropriate output (estimates and the precision thereof) is stored in output data sets, typically using the generic ODS statement. In most cases, it is also advisable to save parameter names along with their values, to facilitate proper matching between the components of a parameter vector and that of the corresponding variance-covariance matrix.

### **PROC MIANALYZE**

Finally, PROC MIANALYZE combines the  $M$  inferences into a single one, by making use of the theory laid out in Section 7.8.2. Appropriate output data sets generated by the analysis procedure and containing parameter estimates, precision estimates, and parameter names, are used as input data sets to PROC MIANALYZE. Depending on the input procedure, such information is passed on using one or more of the following options: `parms=`, `data=`, `parminfo=`, `covb=`, and/or `xpxi=`.

The parameters to be analyzed are passed on via the `MODELEFFECTS` statement. If some effects correspond to categorical variables, the `CLASS` statement should be used. Unless a dedicated data structure is used to pass on parameter estimates and the corresponding variance-covariance matrices, the `STDERR` statement needs to be used to pass on the corresponding standard errors. In that case, both estimates and standard errors come from an ordinary SAS data set.

The `TEST` statement allows testing of hypotheses about linear combinations of the parameters. The statement is based on Rubin (1987), and uses a  $t$  distribution which is the univariate version of the work by Li, Raghunathan, and Rubin (1991), described in Section 7.8.3. Several tests can be combined; each hypothesis testing can be simple or compound.

## **7.8.7 SAS Code for Age-related Macular Degeneration Trial**

The three steps associated with MI, discussed earlier in this section, will be illustrated using the ARMD analysis reported in Section 7.8.5.

### **The MI Procedure for the Imputation Task**

PROC MI is used to generate the imputations. It creates  $M$  imputed data sets from an input data set, physically stored in a single data set with indicator variable `_imputation_`, created by the procedure, to separate the imputed copies.

For imputations from a multivariate Gaussian imputation model, the following MI program can be used:

#### **PROGRAM 7.18 PROC MI for the Imputation Task, using MCMC**

```
proc mi data=armd13 seed=486048 out=armd13a simple n impute=10 round=0.1;
var lesion diff4 diff12 diff24 diff52;
by treat;
run;
```

We have chosen to generate  $M = 10$  imputed data sets, rather than the default number of 25. Imputed values are rounded to one decimal place (by including the `round=` option). Simple statistics are displayed in the output, because of the inclusion of the `simple` option. As stated before, we allow for the imputation model to depend on `lesion` (treated as a continuous variable) and `treat` (treated as categorical, by including it in the `BY` statement).

The `seed=` option is useful when a given data analysis needs to be reproducible, because then every time the same seed is used on the same input data and with the same imputation model (and hence the same program), the same random values will be generated.

Observe that no imputation method is specified, i.e., that there is no MONOTONE, MCMC, or FCS statement. As a result, the default MCMC method will be invoked. To use, for example, FCS, Program 7.19 can be used instead.

### **PROGRAM 7.19 PROC MI for the Imputation Task, using FCS**

```
proc mi data=m.armd13 seed=486048 simple out=m.armd13fcs nimpute=30
      round=0.01;
fcs reg(diff4=lesion);
fcs reg(diff12=lesion diff4);
fcs reg(diff24=lesion diff4 diff12);
fcs reg(diff52=lesion diff4 diff12 diff24);
var lesion diff4 diff12 diff24 diff52;
by treat;
run;
```

Note that, after carrying out the imputation step, the data are still in horizontal format and need to put in the longitudinal, or vertical, format again, which will be done at the outset of the analysis step.

#### **The Analysis Task**

The imputed data sets are now analyzed using a standard complete data procedure. It is important to include `BY _imputation_` to ensure that a separate analysis be carried out for each completed data set.

Also, parameter estimates and their estimated covariance matrices need to be stored in appropriate output data sets, so they can be passed on to the MIANALYZE procedure. We will return to this when discussing the inference step.

To prepare for the data analysis, indicator variables are created, and then the data are sorted by imputation number. A step, specific for our analysis, is that we need to dichotomize the variables.

### **PROGRAM 7.20 Dichotomization of imputed data**

```
proc sort data=m.armd13a;
by _imputation_ subject;
run;

data m.armd13a;
set m.armd13a;
bindif4=0; if diff4 <= 0 then bindif4=1;
bindif12=0; if diff12 <= 0 then bindif12=1;
bindif24=0; if diff24 <= 0 then bindif24=1;
bindif52=0; if diff52 <= 0 then bindif52=1;
if diff4=. then bindif4=.;
if diff12=. then bindif12=.;
if diff24=. then bindif24=.;
if diff52=. then bindif52=.;
run;
```

Next, the data are transformed from the horizontal format to a vertical one, to allow for longitudinal analyses.

**PROGRAM 7.21 Transforming a horizontal data set in a vertical data set**

```

data m.armd13b;
set m.armd13a;
array x (4) bindif4 bindif12 bindif24 bindif52;
array y (4) diff4 diff12 diff24 diff52;
do j=1 to 4;
  bindif=x(j);
  diff=y(j);
  time=j;
  output;
end;
run;

```

While the MIXED, GEE, GENMOD, and GLIMMIX procedures can handle CLASS variables, dummies need to be expressly created for use with the NLMIXED procedure.

**PROGRAM 7.22 Creating dummies**

```

data m.armd13c;
set m.armd13b;
time1=0;
time2=0;
time3=0;
time4=0;
trttime1=0;
trttime2=0;
trttime3=0;
trttime4=0;
if time=1 then time1=1;
if time=2 then time2=1;
if time=3 then time3=1;
if time=4 then time4=1;
if (time=1 & treat=1) then trttime1=1;
if (time=2 & treat=1) then trttime2=1;
if (time=3 & treat=1) then trttime3=1;
if (time=4 & treat=1) then trttime4=1;
run;

proc sort data=m.armd13cs;
by _imputation_ subject time;
run;

```

The GENMOD or GEE procedures can then be called for a GEE analysis.

**PROGRAM 7.23 GEE after multiple imputation**

```

proc gee data=armd13c;
class time subject;
by _imputation_;
model bindif = time1 time2 time3 time4 trttime1 trttime2 trttime3 trttime4
  / noint dist=binomial;
repeated subject=subject / withinsubject=time type=exch modelse covb;
ods output GEEEmpPEst=gmparms parminfo=gmpinfo GEERCov=gmcovb;
run;

```

While we could have used the coding `time treat*time`, the already created dummies are used instead. This makes no difference. The BY statement has been added, as well as the ODS statement, to store the parameter estimates and the covariance parameters. For the latter, the `parminfo=` option is used next to the `covb=` option, to ensure that the proper names of the covariate effects are mapped to abbreviations of type `Prm1`, etc. The parameter estimates are generated by default. The output of the GEE procedure will be a GEE analysis for each of the ten imputed data sets. As such, they represent an intermediate step in the full multiple imputation analysis and are of no direct scientific interest. Formal inference needs to be conducted using only the results from the inference step.

Because the `noint` option was included, the effect `Prm1` formally exists when PROC GENMOD is used (not when PROC GEE is used), but is unavailable as a parameter estimate. It is, therefore, necessary to delete it from the parameter information, as in Program 7.24:

#### **PROGRAM 7.24 Deletion of redundant intercept name**

```
data gmpinfo;
set gmpinfo;
if parameter='Prm1' then delete;
run;
```

The above program should not be used with PROC GEE.

Analogously, the GLMM analysis can be conducted on the multiple imputed data sets. Evidently, both the procedures GLIMMIX and NLMIXED can be used. It is interesting to illustrate the use of a programming-type procedure, such as NLMIXED, in conjunction with multiple imputation.

#### **PROGRAM 7.25 GLMM after multiple imputation**

```
proc nlmixed data=armd13c qpoints=20 maxiter=100 technique=newrap cov ecov;
by _imputation_;
eta = beta11*time1+beta12*time2+beta13*time3+beta14*time4+b
      +beta21*trttime1+beta22*trttime2+beta23*trttime3+beta24*trttime4;
p = exp(eta)/(1+exp(eta));
model bindif ~ binary(p);
random b ~ normal(0,tau*tau) subject=subject;
estimate 'tau2' tau*tau;
ods output ParameterEstimates=nlpars
          CovMatParmEst=nlcovb
          AdditionalEstimates=nlparsa
          CovMatAddEst=nlcovba;
run;
```

Apart from adding the BY statement, we now also generate four output data sets using the ODS statement. For the standard model parameters, we only need the `parameterestimates=` and `covmatparmest=` options. If, in addition, multiple imputation inference is requested about additional estimates, then they can be saved as well using the `additionalestimates=` and `covmataddest=` options. However, it is also possible to calculate the additional estimates directly from the results of the inference step, i.e., to conduct multiple imputation inference first and then calculate additional estimates, rather than the other way around. For both covariance matrices to be generated, the options `cov` and `ecov`, respectively, need to be included into the PROC NLMIXED statement.

For both the GEE and GLMM models, we can now conduct multiple imputation inference, following Rubin's combination rules.

### The Inference Task

Applying the MIANALYZE procedure to the GEE analysis on the ARMD data, presented in Section 7.8.7, can be done using the code in Program 7.26.

#### **PROGRAM 7.26 Inference step after GEE**

```
proc mianalyze parms=gmparms covb=gmcovb parminfo=gmpinfo wcov bcov tcov;
model effects time1 time2 time3 time4 trttime1 trttime2 trttime3 trttime4;
run;
```

Conducting multiple imputation inference for the NLMIXED analysis, presented in Section 7.8.7, is done by means of Program 7.27.

#### **PROGRAM 7.27 Inference step after GLMM**

```
proc mianalyze parms=nlparms covb=nlcovb wcov bcov tcov;
model effects beta11 beta12 beta13 beta14 beta21 beta22 beta23 beta24;
run;
```

### 7.8.8 Creating Monotone Missingness

When missingness is non-monotone, we might think of several mechanisms operating simultaneously: e.g., a simple (MCAR or MAR) mechanism for the intermediate missing values and a more complex (MNAR) mechanism for the missing data past the moment of dropout. However, analyzing such data is complicated since many model strategies, especially those under the assumption of MNAR, have been developed for dropout only. Therefore, a solution might be to generate multiple imputations that make all patterns monotone, by use of Program 7.28.

#### **PROGRAM 7.28 Creating monotone missingness**

```
mcmc impute=monotone;
```

Once done, we can apply a method of choice to the so-completed multiple sets of data. Note that this is different from the monotone method in PROC MI, intended to fully complete already monotone sets of data.

## 7.9 An Overview of Sensitivity Analysis

---

All methods considered so far are valid under MAR and then evidently also under MCAR. The only exception is unweighted GEE, for which in general MCAR is required.

We should not lose sight of the fact that an MNAR mechanism might be operating while it is at the same time formally impossible to distinguish between MAR and MNAR mechanisms, based on observed data alone (Molenberghs et al., 2008). Thus, while it is formally possible to fit models under the assumption of MNAR (Diggle and Kenward, 1994; Verbeke and Molenberghs, 2000, Ch. 18), these should not be considered as evidence for or against MNAR. It is a more viable route to explore how sensitive key inferences (e.g., in terms of parameter estimation or hypothesis testing) are to varying assumptions about the missing-data mechanism. This type of sensitivity to non-identifiable assumptions has been reported in various publications (Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Molenberghs and Kenward, 2007).

Therefore, a sensible compromise between blindly shifting to MNAR models or ignoring them altogether, is to make them a component of a sensitivity analysis.

Broadly, we could define a sensitivity analysis as one in which several statistical models are considered simultaneously and/or where a statistical model is further scrutinized using specialized tools (such as diagnostic measures). This rather loose and very general definition encompasses a wide variety of approaches. The simplest procedure is to fit a selected number of (MNAR) models that are all deemed plausible, or one in which a preferred (primary) analysis is supplemented with a number of variations. The extent to which conclusions (inferences) are stable across such ranges provides an indication about the belief that can be put into them. Variations to a basic model can be constructed in different ways. The most obvious strategy is to consider various dependencies of the missing data process on the outcomes and/or on covariates. Alternatively, the distributional assumptions of the models can be changed. For example, it is natural to start from a primary model of MAR type, and then to consider variations of an MNAR nature.

Several authors have proposed the use of global and local influence tools (Verbeke et al., 2001; Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005). An important question is to what exactly are the sources causing an MNAR model to provide evidence for MNAR against MAR? There is evidence to believe that a multitude of outlying aspects, but not necessarily the (outlying) nature of the missingness mechanism in one or a few subjects, is responsible for an apparent MNAR mechanism (Jansen et al., 2006). The consequence of this is that local influence should be applied and interpreted with due caution. This methodology will be illustrated in Section 7.10.

Another route for sensitivity analysis is by making use of pattern-mixture models (Little, 1993, 1994; Thijs et al., 2002; Michiels et al., 2002). In a PMM, the joint distribution of  $\mathbf{Y}_i$  and  $\mathbf{R}_i$  is factored as the conditional distribution of  $\mathbf{Y}_i$  given  $\mathbf{R}_i$  and the marginal distribution of  $\mathbf{R}_i$ . Recently, this family has gained considerable interest, also due to the work of Carpenter, Roger, and Kenward (2013) and Carpenter and Kenward (2013). Some PMM-based strategies are implemented in the SAS procedure MI, through the MNAR statement. This family will be examined in detail in Section 7.11.

A further framework consists of so-called shared parameter models, where random effects are employed to describe the relationship between the measurement and dropout processes (Wu and Carroll, 1988; DeGruttola and Tu, 1994).

Robins, Rotnitzky, and Scharfstein (1998) discuss sensitivity analysis in a semi-parametric context.

Further, within the selection model framework, Baker, Rosenberger, and DerSimonian (1992) proposed a model for multivariate and longitudinal binary data, subject to non-monotone missingness. Jansen et al. (2003) extended this model to allow for (possibly continuous) covariates, and developed a local influence strategy.

Finally, classical inference procedures account for the imprecision resulting from the stochastic component of the model. Less attention is devoted to the uncertainty arising from (unplanned) incompleteness in the data, even though the majority of clinical studies suffer from incomplete follow-up. Molenberghs et al. (2001) acknowledge both the status of imprecision, due to (finite) random sampling, as well as ignorance, due to incompleteness. Further, both can be combined into uncertainty (Kenward, Molenberghs, and Goetghebeur, 2001).

## 7.10 Sensitivity Analysis Using Local Influence

---

We first introduce the Diggle and Kenward (1994) model, combining a linear mixed model with a model for dropout based on logistic regression, in the spirit of (7.3.11). Thereafter, the use of local influence to examine sensitivity is described.

### 7.10.1 The Model of Diggle and Kenward (DK; 1994)

In agreement with notation introduced in Section 7.3, we assume that a vector of outcomes  $\mathbf{Y}_i$  is designed to be measured. If dropout occurs,  $\mathbf{Y}_i$  is only partially observed. We denote the occasion at which dropout occurs by  $D_i > 1$ , and  $\mathbf{Y}_i$  is split into the  $(D_i - 1)$ -dimensional observed component  $\mathbf{Y}_i^o$  and the  $(n_i - D_i + 1)$ -dimensional missing component  $\mathbf{Y}_i^m$ . In case of no dropout, we let  $D_i = n_i + 1$ , and  $\mathbf{Y}_i$  equals  $\mathbf{Y}_i^o$ . The likelihood contribution of the  $i$ th subject, based on the observed data  $(\mathbf{y}_i^o, d_i)$ , is proportional to the marginal density function

$$\begin{aligned} f(\mathbf{y}_i^o, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\mathbf{y}_i, d_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d\mathbf{y}_i^m \\ &= \int f(\mathbf{y}_i | \boldsymbol{\theta}) f(d_i | \mathbf{y}_i, \boldsymbol{\psi}) d\mathbf{y}_i^m, \end{aligned} \quad (7.10.30)$$

in which a marginal model for  $\mathbf{Y}_i$  is combined with a model for the dropout process, conditional on the response, and where  $\boldsymbol{\theta}$  and  $\boldsymbol{\psi}$  are vectors of unknown parameters in the measurement model and dropout model, respectively.

Let  $\mathbf{h}_{ij} = (y_{i1}, \dots, y_{ij-1})$  denote the observed history of subject  $i$  up to time  $t_{i,j-1}$ . The Diggle-Kenward model for the dropout process allows the conditional probability for dropout at occasion  $j$ , given that the subject was still observed at the previous occasion, to depend on the history  $\mathbf{h}_{ij}$  and the possibly unobserved current outcome  $y_{ij}$ , but not on future outcomes  $y_{ik}$ ,  $k > j$ . These conditional probabilities  $P(D_i = j | D_i \geq j, \mathbf{h}_{ij}, y_{ij}, \boldsymbol{\psi})$  can now be used to calculate the probability of dropout at each occasion:

$$\begin{aligned} P(D_i = j | \mathbf{y}_i, \boldsymbol{\psi}) &= P(D_i = j | \mathbf{h}_{ij}, y_{ij}, \boldsymbol{\psi}) \\ &= \begin{cases} P(D_i = j | D_i \geq j, \mathbf{h}_{ij}, y_{ij}, \boldsymbol{\psi}) & j = 2, \\ P(D_i = j | D_i \geq j, \mathbf{h}_{ij}, y_{ij}, \boldsymbol{\psi}) \\ \times \prod_{k=2}^{j-1} [1 - P(D_i = k | D_i \geq k, \mathbf{h}_{ik}, y_{ik}, \boldsymbol{\psi})] & j = 3, \dots, n_i, \\ \prod_{k=2}^{n_i} [1 - P(D_i = k | D_i \geq k, \mathbf{h}_{ik}, y_{ik}, \boldsymbol{\psi})] & j = n_i + 1. \end{cases} \end{aligned}$$

Diggle and Kenward (1994) combine a multivariate normal model for the measurement process with a logistic regression model for the dropout process. More specifically, the measurement model assumes that the vector  $\mathbf{Y}_i$  of repeated measurements for the  $i$ th subject satisfies the linear regression model  $\mathbf{Y}_i \sim N(X_i \boldsymbol{\beta}, V_i)$ , ( $i = 1, \dots, N$ ). The matrix  $V_i$  can be left unstructured or assumed to be of a specific form, e.g., resulting from a linear mixed model, a factor-analytic structure, or spatial covariance structure (Verbeke and Molenberghs, 2000).

In the particular case that a linear mixed model is assumed, we write

$$\mathbf{Y}_i = X_i \boldsymbol{\beta} + Z_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (7.10.31)$$

(Verbeke and Molenberghs, 2000) where  $\mathbf{Y}_i$  is the  $n$  dimensional response vector for subject  $i$ ,  $1 \leq i \leq N$ ;  $N$  is the number of subjects;  $X_i$  and  $Z_i$  are  $(n \times p)$  and  $(n \times q)$  known design matrices;  $\boldsymbol{\beta}$  is the  $p$  dimensional vector containing the fixed effects; and  $\mathbf{b}_i \sim N(\mathbf{0}, G)$  is the  $q$  dimensional vector containing the random effects. The residual components  $\boldsymbol{\varepsilon}_i \sim N(0, \Sigma_i)$ .

The logistic dropout model can, for example, take the form:

$$\begin{aligned} \text{logit } [P(D_i = j \mid D_i \geq j, \mathbf{h}_{ij}, y_{ij}, \boldsymbol{\psi})] \\ = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}. \end{aligned} \quad (7.10.32)$$

More general models can easily be constructed by including the complete history  $\mathbf{h}_{ij} = (y_{i1}, \dots, y_{i;j-1})$ , as well as external covariates, in the above conditional dropout model. Note also that, strictly speaking, we could allow dropout at a specific occasion to be related to all future responses as well. However, this is rather counter-intuitive in many cases. Moreover, including future outcomes seriously complicates the calculations since computation of the likelihood (7.10.30) then requires evaluation of a possibly high-dimensional integral. Note also that special cases of model (7.10.32) are obtained from setting  $\psi_2 = 0$  or  $\psi_1 = \psi_2 = 0$ , respectively. In the first case, dropout is no longer allowed to depend on the current measurement, implying MAR. In the second case, dropout is independent of the outcome, which corresponds to MCAR.

Diggle and Kenward (1994) obtained parameter and precision estimates by maximum likelihood. The likelihood involves marginalization over the unobserved outcomes  $\mathbf{Y}_i^m$ . Practically, this involves relatively tedious and computationally demanding forms of numerical integration. This, combined with likelihood surfaces tending to be rather flat, makes the model difficult to use. These issues are related to the problems to be discussed next.

### 7.10.2 Local Influence

The local influence approach, suggested by Cook (1986), can be used to investigate the effect of extending an MAR model for dropout in the direction of MNAR dropout (Verbeke et al., 2001).

We start from the DK model introduced in Section 7.10.1. Since no data would be observed otherwise, we assume that the first measurement  $Y_{i1}$  is obtained for every subject in the study. We denote the probability of dropout at occasion  $k$ , given that the subject was still in the study up to occasion  $k$  by  $g(\mathbf{h}_{ik}, y_{ik})$ . For the dropout process, we now consider an extension of model (7.10.32), which can be written as

$$\begin{aligned} \text{logit } [g(\mathbf{h}_{ik}, y_{ik})] &= \text{logit } [P(D_i = k \mid D_i \geq k, \mathbf{y}_i)] \\ &= \mathbf{h}_{ik} \boldsymbol{\psi} + \omega y_{ik}. \end{aligned} \quad (7.10.33)$$

When  $\omega$  equals zero and the model assumptions made are correct, the posited dropout model is MAR, and all parameters can be estimated using standard software since the measurement and dropout model can then be fitted separately. If  $\omega \neq 0$ , the dropout process is assumed to be MNAR. Now, a dropout model might be found to be MNAR solely because one or a few influential subjects have driven the analysis. To investigate sensitivity of estimation of quantities of interest, such as treatment effect, growth parameters, or the dropout model parameters, with respect to assumptions about the dropout model, we consider the following perturbed version of (7.10.33):

$$\begin{aligned} \text{logit } [g(\mathbf{h}_{ik}, y_{ik})] &= \text{logit } [P(D_i = k \mid D_i \geq k, \mathbf{y}_i, W_i)] \\ &= \mathbf{h}_{ik} \boldsymbol{\psi} + \omega_i y_{ik} \quad i = 1, \dots, N. \end{aligned} \quad (7.10.34)$$

There is a fundamental difference with model (7.10.33) since the  $\omega_i$  should not be viewed as parameters: They are local, individual-specific perturbations around a

null model. In our case, the null model will be the MAR model, corresponding to setting  $\omega = 0$  in (7.10.33). Thus, the  $\omega_i$  are perturbations that will be used only to derive influence measures (Cook, 1986).

This scheme enables studying the effect of how small perturbations in the MNAR direction can have a large impact on key features of the model. Practically, one way of doing this is to construct local influence measures (Cook, 1986). Clearly, not all possible forms of impact resulting from sensitivity to dropout model assumptions, will be found in this way, and the method proposed here should be viewed as one component of a sensitivity analysis (e.g., Molenberghs, Kenward, and Goetghebeur, 2001).

When small perturbations in a specific  $\omega_i$  lead to relatively large differences in the model parameters, it suggests that the subject is likely to drive the conclusions.

Cook (1986) suggests that more confidence can be put in a model that is relatively stable under small modifications. The best known perturbation schemes are based on case deletion (Cook and Weisberg, 1982) in which the effect is studied of completely removing cases from the analysis. A quite different paradigm is the local influence approach where we investigate how the results of an analysis are changed under small perturbations of the model. In the framework of the linear mixed model, Beckman, Nachtsheim, and Cook (1987) used local influence to assess the effect of perturbing the error variances, the random-effects variances, and the response vector. In the same context, Lesaffre and Verbeke (1998) have shown that the local influence approach is also useful for the detection of influential subjects in a longitudinal data analysis. Moreover, since the resulting influence diagnostics can be expressed analytically, they often can be decomposed in interpretable components, which yield additional insights into the reasons why some subjects are more influential than others.

We are interested in the influence of MNAR dropout on the parameters of interest. This can be done in a meaningful way by considering (7.10.34) as the dropout model. Indeed,  $\omega_i = 0$  for all  $i$  corresponds to an MAR process, which cannot influence the measurement model parameters. When small perturbations in a specific  $\omega_i$  lead to relatively large differences in the model parameters, this suggests that these subjects might have a large impact on the final analysis. However, even though we might be tempted to conclude that such subjects drop out non-randomly, this conclusion is misguided because we are not aiming to detect (groups of) subjects that drop out non-randomly but rather subjects that have a considerable impact on the dropout and measurement model parameters. Indeed, a key observation is that a subject that drives the conclusions towards MNAR might be doing so, not only because its true data generating mechanism is of an MNAR type, but also for a wide variety of other reasons, such as an unusual mean profile or autocorrelation structure. Earlier analyses have shown that this might indeed be the case. Likewise, it is possible that subjects, deviating from the bulk of the data because they are generated under MNAR, go undetected by this technique.

Let us now introduce the key concepts of local influence. We denote the log-likelihood function corresponding to model (7.10.34) by

$$\ell(\gamma|\omega) = \sum_{i=1}^N \ell_i(\gamma|\omega_i),$$

in which  $\ell_i(\gamma|\omega_i)$  is the contribution of the  $i$ th individual to the log-likelihood, and where  $\gamma = (\theta, \psi)$  is the  $s$ -dimensional vector, grouping the parameters of the measurement model and the dropout model, not including the  $N \times 1$  vector  $\omega = (\omega_1, \omega_2, \dots, \omega_N)'$  of weights defining the perturbation of the MAR model. It is assumed that  $\omega$  belongs to an open subset  $\Omega$  of  $\mathbb{R}^N$ . For  $\omega$  equal to  $\omega_0 = (0, 0, \dots, 0)'$ ,  $\ell(\gamma|\omega_0)$  is the log-likelihood function that corresponds to a MAR dropout model.

Let  $\hat{\gamma}$  be the maximum likelihood estimator for  $\gamma$ , obtained by maximizing  $\ell(\gamma|\omega_0)$ , and let  $\hat{\gamma}_\omega$  denote the maximum likelihood estimator for  $\gamma$  under  $\ell(\gamma|\omega)$ . The local influence approach now compares  $\hat{\gamma}_\omega$  with  $\hat{\gamma}$ . Similar estimates indicate that the parameter estimates are robust with respect to perturbations of the MAR model in the direction of non-random dropout. Strongly different estimates suggest that the estimation procedure is highly sensitive to such perturbations, which, in turn, suggests that the choice between an MAR model and a non-random dropout model highly affects the results of the analysis. Cook (1986) proposed to measure the distance between  $\hat{\gamma}_\omega$  and  $\hat{\gamma}$  by the so-called likelihood displacement, defined by

$$LD(\omega) = 2[\ell(\hat{\gamma}|\omega_0) - \ell(\hat{\gamma}_\omega|\omega_0)].$$

This takes into account the variability of  $\hat{\gamma}$ . Indeed,  $LD(\omega)$  will be large if  $\ell(\gamma|\omega_0)$  is strongly curved at  $\hat{\gamma}$ , which means that  $\gamma$  is estimated with high precision, and small otherwise. Therefore, a graph of  $LD(\omega)$  versus  $\omega$  contains essential information on the influence of perturbations. It is useful to view this graph as the geometric surface formed by the values of the  $N + 1$  dimensional vector  $\xi(\omega) = (\omega', LD(\omega))'$  as  $\omega$  varies throughout  $\Omega$ .

Since this influence graph can only be depicted when  $N = 2$ , Cook (1986) proposed to look at local influence, i.e., at the normal curvatures  $C_h$  of  $\xi(\omega)$  in  $\omega_0$ , in the direction of some  $N$  dimensional vector  $h$  of unit length. Let  $\Delta_i$  be the  $s$  dimensional vector defined by

$$\Delta_i = \left. \frac{\partial^2 \ell_i(\gamma|\omega_i)}{\partial \omega_i \partial \gamma} \right|_{\gamma=\hat{\gamma}, \omega_i=0}$$

and define  $\Delta$  as the  $(s \times N)$  matrix with  $\Delta_i$  as its  $i$ th column. Further, let  $\ddot{L}$  denote the  $(s \times s)$  matrix of second-order derivatives of  $\ell(\gamma|\omega_0)$  with respect to  $\gamma$ , also evaluated at  $\gamma = \hat{\gamma}$ . Cook (1986) has then shown that  $C_h$  can be easily calculated by

$$C_h = 2|h' \Delta' \ddot{L}^{-1} \Delta h|.$$

Obviously,  $C_h$  can be calculated for any direction  $h$ . One evident choice is the vector  $h_i$  containing one in the  $i$ th position and zero elsewhere, corresponding to the perturbation of the  $i$ th weight only. This reflects the influence of allowing the  $i$ th subject to drop out non-randomly, while the others can only drop out at random. The corresponding local influence measure, denoted by  $C_i$ , then becomes  $C_i = 2|\Delta_i' \ddot{L}^{-1} \Delta_i|$ . Another important direction is the direction  $h_{\max}$  of maximal normal curvature  $C_{\max}$ . It shows how to perturb the MAR model to obtain the largest local changes in the likelihood displacement. It is readily seen that  $C_{\max}$  is the largest eigenvalue of  $-2 \Delta' \ddot{L}^{-1} \Delta$ , and that  $h_{\max}$  is the corresponding eigenvector.

#### EXAMPLE: Age-related macular degeneration trial

In this section, in line with Beunckens et al. (2007) and Molenberghs and Kenward (2007), the visual acuity in the ARMD trial is first analyzed using the DK model. Apart from modeling the three missing data mechanisms MCAR, MAR, and MNAR, explicitly, an ignorable analysis is also conducted. For the measurement model, again, the linear mixed model was used, assuming different intercepts and treatment effects for each of the four time points, with an unstructured covariance matrix, as in (7.5.16). In the full selection models, the dropout is modeled as in (7.10.32). Parameter estimates and corresponding standard errors of the fixed effects of the measurement model and of the dropout model parameters are given in Table 7.10.

As expected, the parameter estimates and standard errors coincide for the ignorable likelihood analysis and the selection models under MCAR and MAR, except for some negligible numerical noise.

**TABLE 7.10** The Age-related Macular Degeneration Trial. Parameter estimates (standard errors) assuming ignorability, as well as explicitly modeling the missing data mechanism under MCAR, MAR, and MNAR assumptions, for all data.

Effect	Parameter	Ignorable	MCAR	MAR	MNAR
Measurement model					
Int. 4	$\beta_{11}$	54.00 (1.47)	54.00 (1.46)	54.00 (1.47)	54.00 (1.47)
Int.12	$\beta_{21}$	53.01 (1.60)	53.01 (1.59)	53.01 (1.60)	52.98 (1.60)
Int.24	$\beta_{31}$	49.20 (1.74)	49.20 (1.73)	49.19 (1.74)	49.06 (1.74)
Int.52	$\beta_{41}$	43.99 (1.79)	43.99 (1.78)	43.99 (1.79)	43.52 (1.82)
Trt. 4	$\beta_{12}$	-3.11 (2.10)	-3.11 (2.07)	-3.11 (2.09)	-3.11 (2.10)
Trt. 12	$\beta_{22}$	-4.54 (2.29)	-4.54 (2.25)	-4.54 (2.29)	-4.67 (2.29)
Trt. 24	$\beta_{32}$	-3.60 (2.49)	-3.60 (2.46)	-3.60 (2.50)	-3.80 (2.50)
Trt. 52	$\beta_{42}$	-5.18 (2.59)	-5.18 (2.57)	-5.18 (2.62)	-5.71 (2.63)
Dropout model					
Int.	$\psi_0$		-2.79 (0.17)	-1.86 (0.46)	-1.81 (0.47)
Previous	$\psi_1$			-0.020 (0.009)	0.016 (0.022)
Current	$\psi_2$				-0.042 (0.023)
-2 log-likelihood					
		6488.7	6782.7	6778.4	6775.9
Treatment effect at 1 year ( <i>p</i> -value)					
		0.046	0.044	0.048	0.030

Given that the main interest lies in the treatment effect at one year, the corresponding *p*-values are displayed in Table 7.10. In all four cases, this treatment effect is significant.

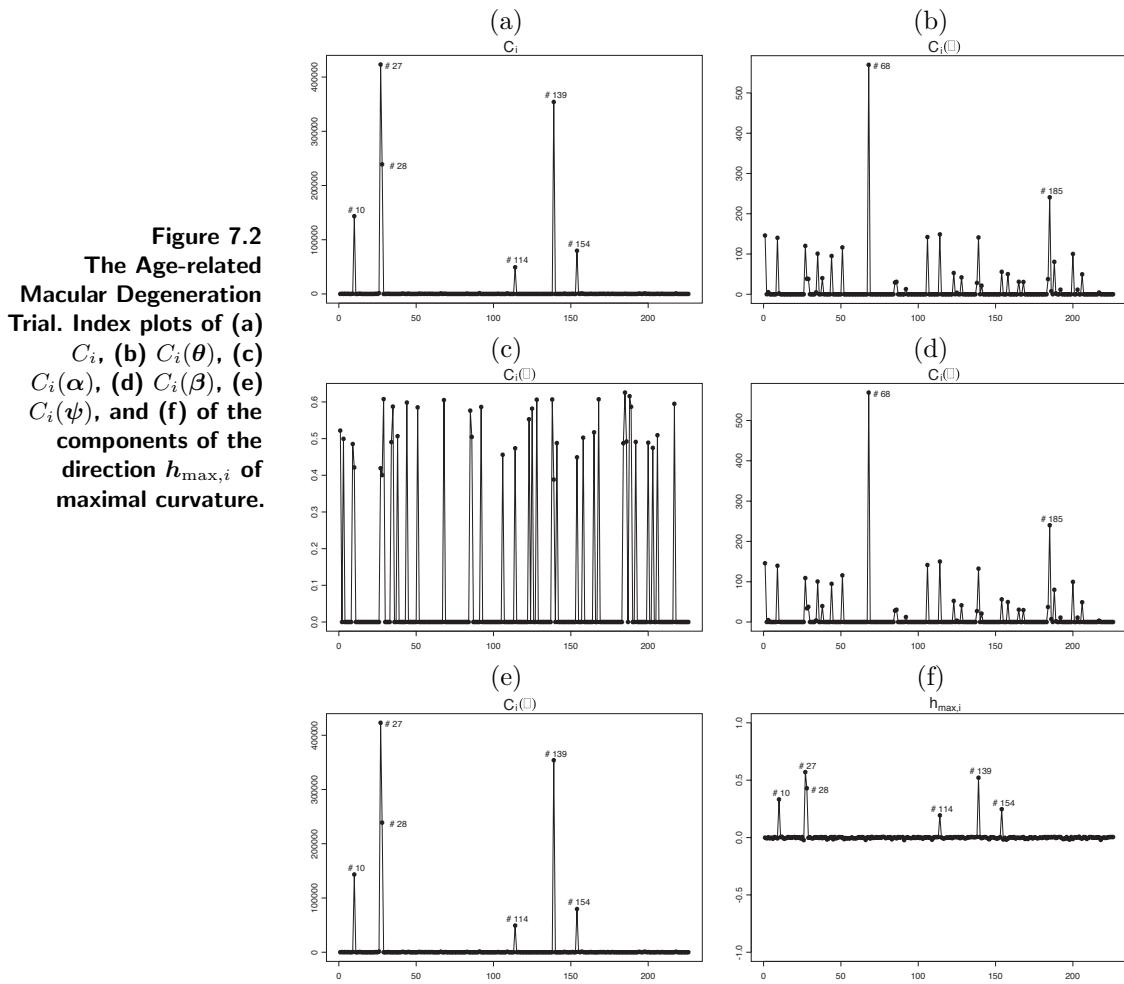
Note that for the MNAR analysis, the estimates of the  $\psi_1$  and  $\psi_2$  parameters are more or less of the same magnitude, but with a different sign. This is in line with the argument of Molenberghs et al. (2001), stating that the dropout often depends on the increment  $y_{ij} - y_{i,j-1}$ . By rewriting the fitted dropout model in terms of the increment,

$$\text{logit} [\text{pr}(D_i = j | D_i \geq j, \mathbf{y}_i)] = -1.81 - 0.026y_{i,j-1} - 0.042(y_{ij} - y_{i,j-1}),$$

we find that the probability of dropout increases with larger negative increments; that is, those patients who showed or would have shown a greater decrease in visual acuity from the previous visit are more likely to drop out.

Turning to local influence. Figure 7.2 displays overall  $C_i$  and influences for sub-vectors  $\boldsymbol{\theta}$ ,  $\boldsymbol{\beta}$ ,  $\boldsymbol{\alpha}$ , and  $\boldsymbol{\psi}$ . In addition, the direction  $\mathbf{h}_{\max}$ , corresponding to maximal local influence, is given. The main emphasis should be put on the relative magnitudes. We observe that patients #10, #27, #28, #114, #139, and #154 have larger  $C_i$  values compared to other patients, which means they can be considered influential. Virtually the same picture holds for  $C_i(\boldsymbol{\psi})$ .

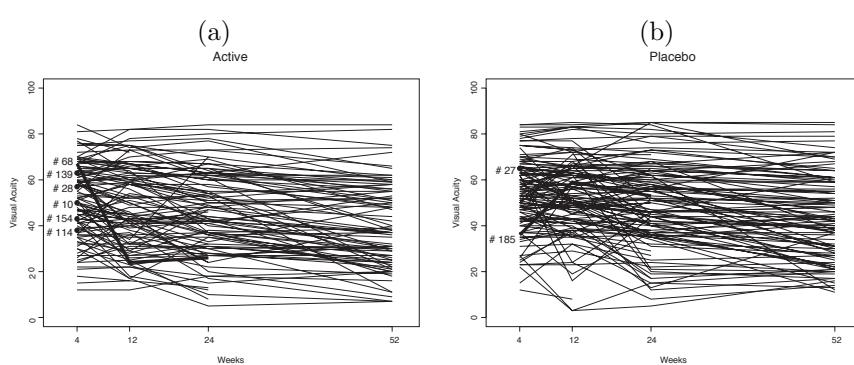
Turning attention now to the influence on the measurement model, we see that for  $C_i(\boldsymbol{\beta})$ , there are no strikingly high peaks, whereas  $C_i(\boldsymbol{\alpha})$  reveals considerable peaks for patients #68 and #185. Note that both patients fail to have a high peak for the overall  $C_i$ , owing to the fact that the scale for  $C_i(\boldsymbol{\alpha})$  is relatively small compared to the overall  $C_i$ . Nevertheless, these patients can still be considered influential. Finally, the direction of maximum curvature reveals the same six influential patients as the overall  $C_i$ .



**Figure 7.2**  
The Age-related  
Macular Degeneration  
Trial. Index plots of (a)  
 $C_i$ , (b)  $C_i(\theta)$ , (c)  
 $C_i(\alpha)$ , (d)  $C_i(\beta)$ , (e)  
 $C_i(\psi)$ , and (f) of the  
components of the  
direction  $h_{\max,i}$  of  
maximal curvature.

In Figure 7.3, the individual profiles of the influential observations are highlighted. Let us take a closer look at these cases. The six patients strongly influencing the dropout model parameters are those dropping out after the first measurement is taken at week 4. All of these patients are in the active treatment arm, except for #27. On the other hand, the two patients with strong influence on the measurement model parameters stay in the study up to week 24 and then have no observation for the last measurement occasion at 1 year. Patient #68 received the active treatment, and his/her visual acuity decreases substantially after week 4, thereafter staying more or less level. Conversely, patient #185 is enrolled in the placebo treatment arm and his/her visual acuity increases after week 4, then sloping downward a little after week 12.

**Figure 7.3**  
The Age-related  
Macular Degeneration  
Trial. Individual profiles  
for both treatment  
arms, with influential  
subjects highlighted.



It is of interest to consider an analysis without these influential observations. Therefore, we applied the selection model on three subsets of the data. The first subset obtains by removing all eight influential patients mentioned before. In the second subset of the data, patients #10, #27, #28, #114, #139, and #154 were removed, since these are overall the most influential ones. Finally, patients #68 and #185, which seemed to be influencing the measurement model the most, were removed, resulting in the third subset. Results of these analyses are shown in Tables 7.11 and 7.12. We compare the results of the MAR and MNAR analyses.

**TABLE 7.11** **The Age-related Macular Degeneration Trial. Parameter estimates (standard errors) explicitly modeling the missing data mechanism under MAR assumptions, after removing the following subsets of subjects Set 1: (10, 27, 28, 114, 139, 154, 68, 185); Set 2: (10, 27, 28, 114, 139, 154); and Set 3: (68, 185).**

Effect	Parameter	Set 1	Set 2	Set 3
		MAR	MAR	MAR
Measurement model				
Int. 4	$\beta_{11}$	54.14(1.51)	54.30(1.47)	53.84(1.48)
Int.12	$\beta_{21}$	53.09(1.64)	53.16(1.59)	52.94(1.60)
Int.24	$\beta_{31}$	49.56(1.77)	49.31(1.74)	49.44(1.73)
Int.52	$\beta_{41}$	44.40(1.82)	44.00(1.79)	44.38(1.78)
Trt. 4	$\beta_{12}$	-3.13(2.17)	-3.28(2.08)	-2.95(2.07)
Trt.12	$\beta_{22}$	-4.48(2.36)	-4.55(2.26)	-4.47(2.26)
Trt.24	$\beta_{32}$	-3.80(2.56)	-3.55(2.48)	-3.85(2.44)
Trt.52	$\beta_{42}$	-5.45(2.66)	-5.06(2.59)	-5.56(2.55)
Dropout model				
Intercept	$\psi_0$	-1.90(0.47)	-1.90(0.47)	-1.85(0.46)
Previous	$\psi_1$	-0.019(0.010)	-0.019(0.010)	-0.020(0.009)
-2 log-likelihood		6535.3	6606.9	6706.4
Treatm. eff. at 1 year ( <i>p</i> -value)		0.040	0.051	0.029

**TABLE 7.12** **The Age-related Macular Degeneration Trial. Parameter estimates (standard errors) explicitly modeling the missing data mechanism under MNAR assumptions, after removing the following subsets of subjects Set 1: (10, 27, 28, 114, 139, 154, 68, 185); Set 2: (10, 27, 28, 114, 139, 154); and Set 3: (68, 185).**

Effect	Parameter	Set 1	Set 2	Set 3
		MNAR	MNAR	MNAR
Measurement model				
Int. 4	$\beta_{11}$	54.15(1.49)	54.30(1.46)	53.84(1.47)
Int.12	$\beta_{21}$	53.06(1.62)	53.13(1.59)	52.91(1.59)
Int.24	$\beta_{31}$	49.46(1.75)	49.20(1.72)	49.31(1.72)
Int.52	$\beta_{41}$	43.97(1.84)	43.58(1.82)	43.90(1.82)
Trt. 4	$\beta_{12}$	-3.13(2.11)	-3.28(2.06)	-2.95(2.05)
Trt.12	$\beta_{22}$	-4.63(2.29)	-4.69(2.24)	-4.60(2.23)
Trt.24	$\beta_{32}$	-4.04(2.49)	-3.79(2.44)	-4.04(2.42)
Trt.52	$\beta_{42}$	-6.12(2.66)	-5.72(2.61)	-6.09(2.58)
Dropout model				
Intercept	$\psi_0$	-1.85(0.49)	-1.85(0.49)	-1.81(0.47)
Previous	$\psi_1$	0.018(0.022)	0.017(0.022)	0.017(0.022)
Current	$\psi_2$	-0.044(0.024)	-0.043(0.024)	-0.043(0.024)
-2 log-likelihood		6532.7	6604.4	6703.8
Treatm. eff. at 1 year ( <i>p</i> -value)		0.021	0.028	0.018

After removing the patients, who have large overall  $C_i$  and  $C_i(\psi)$  values, the estimates of the dropout model parameters  $\psi_1$  and  $\psi_2$  are approximately the same, whereas the estimate of  $\psi_0$  decreases from  $-1.86$  to  $-1.90$  under MAR, and from  $-1.81$  to  $-1.85$  under MNAR. The same can be seen after removing all patients. Considering the treatment effect at 1 year, its estimate under the MAR analysis increases from  $-5.18$  to  $-5.06$ , yielding a slightly increased borderline  $p$ -value, whereas under the MNAR analysis it decreases with  $0.01$ . Together with a decreased standard error this yields a small decrease in the  $p$ -value.

There is no impact on the likelihood ratio test for MAR against MNAR: After removing either patients #10, #27, #28, # 114, #139, and #154, or all influential patients,  $G^2$  remains  $2.5$ . If this likelihood ratio test would follow a standard  $\chi^2_1$ -distribution, we would fail to reject the null hypothesis, which leads us to the MAR assumption. However, the test of MAR against MNAR is non-standard and the conventional chi-squared approximation cannot be used for its null distribution (Rotnitzky et al., 2000, Jansen et al., 2006).

Finally, we perform the same analyses on the third subset, with patients #68 and # 185 removed. Both for the MAR and MNAR analysis, the estimate of the treatment effect at 1 year decreases quite a lot, from  $-5.18$  to  $-5.56$  and from  $-5.71$  to  $-6.09$  respectively. Consequently, the  $p$ -value also drops down from  $0.048$  to  $0.029$  under MAR and from  $0.030$  to  $0.018$  under the MNAR analysis. The deviance for the likelihood ratio test for MAR against MNAR only changes slightly from  $2.5$  to  $2.6$ .

## 7.11 Sensitivity Analysis Based on Multiple Imputation and Pattern-Mixture Models

---

In Section 7.11.1, the strategies to fit pattern-mixture models as described in Molenberghs and Kenward (2007, Ch. 17) are reviewed and applied to the ARMD data. In Section 7.11.2, sensitivity analyses methods, combining pattern-mixture models (PMM) and multiple imputation are described and their SAS implementation discussed.

### 7.11.1 Pattern-Mixture Strategies

PMM are inherently under-identified, because the data available for modeling within a given pattern are by definition confined to the observed components only. Verbeke and Molenberghs (2000) and Molenberghs and Kenward (2007) consider several identification strategies.

**Strategy 1.** Little (1993, 1994) addresses the under-identification through the use of identifying restrictions: Within a given pattern, the predictive distribution of the unobserved measurements, given the observed ones, is set equal to its counterpart from other patterns (e.g., the completers' pattern, termed CCMV; the neighboring pattern, termed NCMV; or a particular combination across all patterns from which the distribution is estimable, termed ACMV).

**Strategy 2.** As an alternative to identifying restrictions, model simplification can be undertaken to identify the parameters. The advantage is that the number of parameters decreases, which is desirable since the length of the parameter vector is a general issue with pattern-mixture models. Hogan and Laird (1997) noted that, to estimate the large number of parameters in general pattern-mixture

models, we have to make the awkward requirement that each dropout pattern occurs sufficiently often. Broadly, we distinguish between two interconnected types of simplifications.

- **Strategy 2a.** Trends can be restricted to functional forms supported by the information available within a pattern. For example, a linear or quadratic time trend is easily extrapolated beyond the last obtained measurement. We merely need to provide an ad hoc solution for the first or the first few patterns. To fit such models, a conventional model building exercise is conducted within each of the patterns separately.
- **Strategy 2b.** Alternatively, we can choose to let the model parameters vary across patterns in a controlled parametric way. Thus, rather than estimating a separate time trend within each pattern, we might for example assume that the time evolution within a pattern is unstructured, but parallel across patterns. This can be done by treating pattern as a covariate. The available data can be used to assess whether such simplifications are supported over the time ranges for which information is collected.

While the second strategy is computationally the simpler, there is a price to pay. Such simplified models, qualified as “assumption rich” by Sheiner, Beal, and Dunne (1997), are also making untestable assumptions, exactly as in the selection model case. Using the fitted profiles to predict their evolution beyond the time of dropout is nothing but extrapolation. It is possible only by making the models sufficiently simple. It is, for example, not possible to assume an unstructured time trend in incomplete patterns and then still extrapolate in an unambiguous fashion. In contrast, assuming a linear time trend allows estimation in all patterns containing at least two measurements. However, it is less obvious what the precise nature of the dropout mechanism is. Kenward, Molenberghs, and Thijs (2003) examined what restrictions need to be imposed, in the context of longitudinal data with dropout, to ensure that the dropout probability does not depend on future measurements, given past and current values. Strategy 2 is not compliant with this requirement, but the same is true for CCMV and NCMV.

A final observation, applying to both strategies, is that pattern-mixture models do not always automatically provide estimates and standard errors of marginal quantities of interest, such as overall treatment effect or overall time trend. Hogan and Laird (1997) provided a way to derive selection model quantities from the pattern-mixture model. This is a first instance in the PMM context where multiple imputation comes in handy. Several authors have followed this idea to formally compare the conclusions from a selection model with the selection model parameters derived from a pattern-mixture model (Verbeke, Lesaffre, and Spiessens, 1998; Michiels, Molenberghs, and Lipsitz, 1999).

To better see how this method works, we briefly sketch the sequence of steps to be followed.

1. Fit a model to the pattern  $t$ -specific identifiable densities:  $f_t(y_1, \dots, y_t)$ . This results in a parameter estimate,  $\hat{\gamma}_t$ .
2. Select an identification method of choice.
3. Using this identification method, determine the conditional distributions of the unobserved outcomes, given the observed ones:

$$f_t(y_{t+1}, \dots, y_T | y_1, \dots, y_t). \quad (7.11.35)$$

4. Using standard MI methodology, draw multiple imputations for the unobserved components, given the observed outcomes and the correct pattern-specific density (7.11.35).

5. Analyze the multiply-imputed sets of data using the method of choice. This can be another pattern-mixture model, but also a selection model or any other desired model.
6. Inferences can be conducted using the standard combination rules.

**EXAMPLE: Age-related macular degeneration trial**

We now consider the use of pattern-mixture models for these data. Here, we will apply the first strategy making, use of CCMV, NCMV, and ACMV identifying restrictions.

The results for the three types of restrictions are shown in Table 7.13. After applying each one of the three restrictions, the same selection model as before is fitted. It can be seen from the estimates and associated standard errors that there is little difference in conclusions between the strategies.

**TABLE 7.13** The Age-related Macular Degeneration Trial. Parameter estimates (standard errors) and *p*-values resulting from the pattern-mixture model using identifying restrictions ACMV, CCMV, and NCMV.

Effect	Parameter	ACMV	CCMV	NCMV
Parameter estimate (standard error)				
Intercept 4	$\beta_{11}$	54.00(1.47)	54.00(1.47)	54.00(1.47)
Intercept 12	$\beta_{21}$	52.87(1.68)	52.92(1.61)	52.86(1.63)
Intercept 24	$\beta_{31}$	48.65(2.00)	49.16(1.87)	48.77(1.78)
Intercept 52	$\beta_{41}$	44.19(2.14)	44.69(2.54)	44.00(1.80)
Treatment 4	$\beta_{12}$	-3.11(2.10)	-3.11(2.10)	-3.11(2.10)
Treatment 12	$\beta_{22}$	-4.18(2.48)	-4.07(2.30)	-4.40(2.42)
Treatment 24	$\beta_{32}$	-4.36(3.83)	-5.14(3.61)	-4.19(2.62)
Treatment 52	$\beta_{42}$	-5.04(3.86)	-2.33(4.93)	-4.89(2.70)
<i>p</i> -values				
Intercept 4	$\beta_{11}$	---	---	---
Intercept 12	$\beta_{21}$	< .0001	< .0001	< .0001
Intercept 24	$\beta_{31}$	< .0001	< .0001	< .0001
Intercept 52	$\beta_{41}$	< .0001	< .0001	< .0001
Treatment 4	$\beta_{12}$	---	---	---
Treatment 12	$\beta_{22}$	0.092	0.077	0.069
Treatment 24	$\beta_{32}$	0.271	0.173	0.110
Treatment 52	$\beta_{42}$	0.211	0.647	0.071

In the pattern-mixture approach, we use information from different patterns to multiply impute new values whenever the observations are missing. Borrowing information from more distant patterns, such as the complete cases, can introduce extra variability, depending on the nature of the conditional distributions sampled from. It is not unexpected, therefore, for the variability to be smallest when applying NCMV, as seen in the standard errors.

It can be seen from these analyses that the treatment effect at week 52 is not statistically significant, in contrast to the conclusions based on Tables 7.10 and 7.11. The *p*-value is closest to significance with NCMV restrictions.

The fact that no significant treatment effect is found here, suggests caution concerning the conclusions obtained under the selection model formulation. This implies that a significant treatment effect is conditional upon the MAR assumption

holding. We would feel more comfortable about a significant treatment effect if it were holding across MAR and a number of MNAR scenarios. Thus, at best, it is fair to say that there is a weak evidence only for a treatment effect.

### 7.11.2 Pattern-Mixture Based Sensitivity Analysis

The PMM framework is very rich, and sensitivity analyses can be conducted from various perspectives. For example, we can use the PMM framework to examine the impact of certain departures from MAR (e.g., when patients would evolve differently after dropout, while still on the same treatment). A different perspective is the examination of potential outcomes that patients would have had, had they switched to alternative treatment strategies after dropping out (Little and Kang, 2015).

From a technical point of view, the algorithm described on page 372 is very general and allows for a variety of sensitivity analysis routes.

First, by simply varying the identifying restrictions (e.g., by juxtaposing ACMV, CCMV, and NCMV; but there are several others) a sensitivity analysis results. Note that ACMV corresponds to MAR in the PMM framework (Molenberghs et al., 1998); the other two can then be seen as deviations from it.

Second, it is of course possible to identify conditional densities of the form (7.11.35) in other ways than through setting them equal to other data-identified densities, or in ways that deviate from them in a controlled way. This is the route taken by Carpenter, Roger, and Kenward (2009), and discussed in detail in Carpenter and Kenward (2013, Ch. 10).

An advantage of using MI is that imputations can be generated in a PMM framework, with analysis conducted in the same or a different framework. For example, the models reported in Table 7.13 are of a selection model type, but imputations were obviously of a PMM signature.

Possible strategies for generating imputations, partially in line with Carpenter, Roger, and Kenward (2009) are as follows:

- Jump to reference. For example, patients receiving active treatment might be made to “jump” to the control group after dropout.
- After dropout in a given pattern (and perhaps in a given treatment group), subjects might be made to shift with a certain amount, relative to the MAR-based prediction. This amount in itself can be varied, from 0 (typically corresponding to MAR), to a prespecified maximal amount.
- Likewise, they might be made to change slope with a certain amount.

Carpenter and Kenward (2009) describe such strategies in the following generic terms:

1. *Separately for each treatment arm, take all patients' pre-deviation data and---assuming MAR---fit a multivariate normal distribution with unstructured mean (i.e., a separate mean for each of the  $1 + p$  baseline plus post-randomization observation times) and unstructured variance-covariance matrix (i.e., a  $(1+p) \times (1+p)$  covariance matrix), (...).*
2. *Separately for each treatment arm, draw a mean vector and variance-covariance matrix from the posterior distribution.*
3. *For each patient who deviates before the end of the study, use the draws from step 2 to build the joint distribution of their pre- and post-deviation outcome data. Suggested options for constructing this are given below.*

4. For each patient who deviates before the end, use their joint distribution in step 3 to construct their conditional distribution of post-deviation given pre-deviation outcome data. Sample their post-deviation data from this conditional distribution, to create a “completed” data set.
5. Repeated steps 2–4  $M$  times, resulting in  $M$  imputed data sets.
6. Fit the substantive model to each imputed data set, and combine the resulting parameter estimates and standard errors using Rubin’s rules for final inference.

For precision estimation, we might also revert to resampling methods, as proposed by Lu (2014).

A special place is reserved for a so-called *tipping point analysis*. This can be undertaken whenever a continuous deviation from MAR is possible. For example, when in a given pattern for a given treatment group, subjects are systematically shifted by a certain amount, this amount can be changed continuously (or in small increments) until the point where significance of a key hypothesis test changes. If this point is unrealistically far away, then confidence in the primary analysis increases. Of course, there are a multitude of ways in which a given primary analysis can be subjected to a tipping point analysis. Exactly how it is conducted will likely depend on substantive considerations.

**EXAMPLE: Age-related macular degeneration trial**

To illustrate sensitivity analysis by way of MNAR adjustments in the multiple imputation process, we apply a shift to missing values in the treated arm, with magnitudes of 0, 10, 15, and 20, at 4, 12, 24, and 52 weeks, respectively. The results of both GEE and GLMM, without and with this adjustment, are shown in Table 7.14. We expect to see the same estimates for the standard (MAR) analyses

**TABLE 7.14** The Age-related Macular Degeneration Trial. Parameter estimates (standard errors) for GEE and GLMM, comparing MAR versions with MNAR analyses based on shifts, identification using the placebo group only, and NCMV.

Effect	Par.	MAR	shift	placebo	NCMV
Generalized estimating equations					
Int.4	$\beta_{11}$	-0.82(0.20)	-0.73(0.20)	-0.81(0.20)	-0.83(0.21)
Int.12	$\beta_{21}$	-0.97(0.22)	-0.71(0.19)	-0.98(0.22)	-1.06(0.21)
Int.24	$\beta_{31}$	-1.07(0.23)	-0.56(0.19)	-1.05(0.22)	-1.00(0.22)
Int.52	$\beta_{41}$	-1.66(0.27)	-0.82(0.20)	-1.58(0.29)	-1.59(0.27)
Trt.4	$\beta_{12}$	0.17(0.29)	0.07(0.28)	0.17(0.28)	0.17(0.29)
Trt.12	$\beta_{22}$	0.56(0.29)	0.29(0.27)	0.56(0.29)	0.67(0.28)
Trt.24	$\beta_{32}$	0.41(0.30)	-0.10(0.27)	0.39(0.29)	0.34(0.29)
Trt.52	$\beta_{42}$	0.41(0.35)	-0.43(0.30)	0.32(0.35)	0.32(0.35)
Generalized linear mixed models					
Int.4	$\beta_{11}$	-1.46(0.36)	-1.32(0.36)	-1.39(0.35)	-1.42(0.35)
Int.12	$\beta_{21}$	-1.75(0.38)	-1.27(0.35)	-1.67(0.37)	-1.80(0.36)
Int.24	$\beta_{31}$	-1.83(0.38)	-1.01(0.34)	-1.78(0.38)	-1.70(0.38)
Int.52	$\beta_{41}$	-2.71(0.45)	-1.47(0.36)	-2.62(0.46)	-2.64(0.44)
Trt.4	$\beta_{12}$	0.32(0.48)	0.12(0.50)	0.25(0.48)	0.24(0.48)
Trt.12	$\beta_{22}$	0.99(0.49)	0.50(0.48)	0.91(0.48)	1.09(0.47)
Trt.24	$\beta_{32}$	0.67(0.51)	-0.19(0.48)	0.62(0.48)	0.53(0.49)
Trt.52	$\beta_{42}$	0.53(0.57)	-0.74(0.51)	0.45(0.56)	0.45(0.55)
R.I. s.d.	$\tau$	2.21(0.26)	2.28(0.25)	2.17(0.25)	2.16(0.24)
R.I. var.	$\tau^2$	4.90(1.14)	5.21(1.15)	4.72(1.09)	4.66(1.05)

as obtained in Table 7.9. However, there are slight differences because the results in Table 7.9 were based on the MCMC imputation method, whereas here FCS was used. The reason is that the MNAR statement requires either MONOTONE or FCS. Given that the data are slightly non-monotone, FCS is the obvious choice.

Turning to the difference between the MAR analyses and those after applying a shift, we observe reasonably large changes, including a sign change for the treatment effects at 24 and 52 weeks. Under MAR, there was only one marginally significant treatment effect (at 12 weeks), even though it is in favor of placebo. After applying the shift, nothing is nearly significant. It is noteworthy that a negative sign points to an effect in favor of the active treatment. This is not surprising, because by applying the shift, we progressively make the treatment more beneficial.

Two further MNAR-based analyses are conducted, both reported in Table 7.14. In the first case, imputation takes place based on the placebo group only. In other words, after dropout, the conditional distribution of the missing measurements given the observed ones is based on only the control arm. The final analysis, NCMV is applied, meaning that the distribution of a missing measurement given its predecessors is based on the adjacent pattern that contains all of these measurements. This imputation is done for each of the two treatment groups separately.

Unlike with the shift analysis, the SAS implementation for the latter two analyses requires data to be monotonically missing. Therefore, two multiple imputation calls are made. In the first one, the eight subjects with a non-monotone pattern are monotonized, under the assumption of MAR. Ten imputations are generated. These data sets are then used as input for one further MNAR imputation. The result is, evidently, ten fully imputed data sets. Details about the SAS code are presented in Section 7.11.3.

### 7.11.3 SAS and Sensitivity Analysis

The key statement to conduct sensitivity analyses of the type reported above is the MNAR statement. It is important to note that the statement requires either MONOTONE or FCS as imputation strategies. Hence, MCMC is not compatible with this tool.

There are two main strategies to apply the MNAR statement. The first one is that of adjustment, using the `adjust` option. It specifies a subset of the variables present in the VAR statement to which a certain adjustment should be applied. It is also possible to specify a subset of the observations to which the adjustment needs to be applied. For example, we can apply an adjustment at certain measurement occasions, for one of several treatment arms. An example is given in the following program (Program 7.29), which is needed to generate the imputations that lead to the results reported in Table 7.14.

#### PROGRAM 7.29 Sensitivity analysis using PROC MI, shift adjustment

```
proc mi data=m.armd13 seed=486048 simple out=m.armd13as1
    nimpute=10 round=0.1;
    title 'Shift multiple imputation';
    class treat;
    var lesion diff4 diff12 diff24 diff52;
    fcs reg;
    mnar adjust (diff12 / shift=10 adjustobs=(treat='2'));
    mnar adjust (diff24 / shift=15 adjustobs=(treat='2'));
    mnar adjust (diff52 / shift=20 adjustobs=(treat='2'));
    by treat;
run;
```

Note that Program 7.29 replaced Program 7.18. The programs for the analysis and inference steps remain exactly the same. The `shift=` adjustment is but one of

several options. For example, rather than an additive shift, a multiplicative scale adjustment can be made using the `scale=` option.

The other main option is `model`. It can be used to specify for which variables what subgroup of the observations is to be used. Subgroups can be defined in a predefined way, using NCMV or CCMV. Alternatively, subgroups can be defined by way of levels of certain variables. An example of the latter is Program 7.30.

### PROGRAM 7.30 Sensitivity analysis using PROC MI, subgroup adjustment

```
proc mi data=m.armd13 seed=486048 simple out=m.armd13as2 nimpute=10;
title 'Model multiple imputation';
class treat;
var lesion diff4 diff12 diff24 diff52;
fcs reg;
mnar model (diff4 / modelobs= (treat='1'));
mnar model (diff12 / modelobs= (treat='1'));
mnar model (diff24 / modelobs= (treat='1'));
mnar model (diff52 / modelobs= (treat='1'));
run;
```

The method is particularly useful, and popular, when this group is defined as a control treatment group. Such a control-based imputation method is known as *copy reference*. The website [www.missingdata.org.uk](http://www.missingdata.org.uk) contains a suite of SAS macros, for various control-based imputation strategies, written by James Roger.

Note that NCMV and CCMV is available only for monotone data, whereas the third option is available with FCS as well.

Should we want to apply NCMV or CCMV, then we can first monotonize the data using standard imputation, and then apply the desired identifying restrictions, as in the following program:

### PROGRAM 7.31 Sensitivity analysis using PROC MI, NCMV

```
proc mi data=m.armd13 seed=486048 simple out=m.armd13as3 nimpute=10;
title 'Montone imputation';
var lesion diff4 diff12 diff24 diff52;
mcmc impute=monotone;
by treat;
run;

proc mi data=m.armd13as3 seed=486048 simple out=m.armd13as4 nimpute=1;
title 'Model multiple imputation';
var lesion diff4 diff12 diff24 diff52;
monotone reg;
mnar model (diff4 diff12 diff24 diff52 / modelobs=ncmv);
by treat;
run;
```

---

#### Output Observations Used for Imputation Models Under MNAR Assumption

Imputed Variable	Observations
diff4	Nonmissing lesion, diff4; Missing diff12, ..., diff52
diff12	Nonmissing lesion, ..., diff12; Missing diff24, diff52
diff24	Nonmissing lesion, ..., diff24; Missing diff52
diff52	Complete Cases

---

In the first MI call, 10 imputations are generated. The output data set of this call is used as input for the next one, where a single imputation is created. Evidently, this effectively creates  $10 \times 1 = 10$  imputations. As a general rule, when multiple imputations are generated in a sequential fashion, the required number of imputations  $M$  should be generated the first time; in every subsequent call, there should then be a single imputation.

The patterns used in the imputation process (second call) are part of the printout:

## 7.12 Concluding Remarks

---

We have shown that analyzing incomplete (longitudinal) data, both of a Gaussian as well as of a non-Gaussian nature, can easily be done under the relatively relaxed assumption of missingness at random (MAR), using standard statistical software tools. Likelihood-based methods include the linear mixed model (e.g., implemented in the SAS procedure MIXED) and generalized linear mixed models (e.g., implemented in the SAS procedures GLIMMIX and NLMIXED). This is termed direct likelihood or ignorable likelihood. Under the same assumptions, ignorable Bayesian analyses can be conducted (e.g., using the SAS procedure MCMC).

In addition, weighted generalized estimating equations can be used under MAR. Its implementation is straightforward thanks to facilities of the SAS procedure GEE.

Finally, a versatile approach, valid under MAR, is to handle incompleteness by way of multiple imputation, after which standard, complete-data analysis methods can be used. SAS offers procedures MI and MIANALYZE to this effect.

All of this implies that traditionally popular but far more restricted modes of analysis, including complete case (CC) analysis, last observation carried forward (LOCF), or other simple imputation methods, ought to be abandoned, given the highly restrictive assumptions on which they are based.

Of course, general missingness not at random can never be entirely excluded, and we should therefore ideally supplement an MAR-based analysis with a suitable chosen set of sensitivity analyses. This area is still in full development, but thanks to the MNAR statement in PROC MI, an array of sensitivity analysis tools are now also provided within the context of standard SAS procedures.

## 7.13 References

---

- Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). Topics in Modelling of Clustered Data. London: Chapman and Hall.
- Afifi, A. and Elashoff, R. (1966). Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association*, **61**, 595–604.
- Baker, S.G., Rosenberger, W.F., and DerSimonian, R. (1992). Closed-form estimates for missing counts in two-way contingency tables. *Statistics in Medicine*, **11**, 643–657.
- Beckman, R.J., Nachtsheim, C.J., and Cook, R.D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, **29**, 413–426.
- Beunckens, C., Molenberghs, G., Thijs, H., and Verbeke, G. (2007). Incomplete hierarchical data. *Statistical Methods in Medical Research*, **16**, 1–36.
- Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

- Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 1014--1029.
- Carey, V.C., Zeger, S.L., and Diggle, P.J. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**, 517--526.
- Carpenter, J.R. and Kenward, M.G. (2013). *Multiple Imputation and Its Applications*. Chichester: John Wiley & Sons.
- Carpenter, J.R., Kenward, M.G., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society, Series A*, **169**, 571--584.
- Carpenter, J.R., Roger, J.H., and Kenward, M.G. (2013). Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, **23**, 1352--1371.
- Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133--169.
- Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman & Hall.
- DeGruttola, V. and Tu, X.M. (1994). Modelling progression of CD4 lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003--1014.
- Dempster, A.P., Laird, N.M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1--38.
- Dempster, A.P. and Rubin, D.B. (1983). Overview. *Incomplete Data in Sample Surveys, Vol. II: Theory and Annotated Bibliography*, W.G. Madow, I. Olkin, and D.B. Rubin (Eds.). New York: Academic Press, pp. 3--10.
- Diggle, P.J. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49--93.
- Fitzmaurice, G.M., Davidian, M., Verbeke, G., and Molenberghs, G. (2009) *Advances in Longitudinal Data Analysis*. London: CRC/Chapman Hall.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*, Texts in Statistical Science. London: Chapman & Hall.
- Glynn, R.J., Laird, N.M., and Rubin, D.B. (1986). Selection modelling versus mixture modelling with non-ignorable nonresponse. In: *Drawing Inferences from Self Selected Samples*, H. Wainer (Ed.). New York: Springer-Verlag, pp. 115--142.
- Hartley, H.O. and Hocking, R. (1971). The analysis of incomplete data. *Biometrics*, **27**, 7783--808.
- Heitjan, F. and Little, R.J.A. (1991). Multiple imputation for the fatal accident reporting system. *Applied Statistics*, **40**, 13-29.
- Hogan, J.W. and Laird, N.M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239--258.
- Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., and Van Steen, K. (2003). A local influence approach applied to binary data from a psychiatric study. *Biometrics*, **59**, 409--418.
- Jansen, I., Hens, N., Molenberghs, G., Aerts, M., Verbeke, G., and Kenward, M.G. (2006). The nature of sensitivity in missing not at random models. *Computational Statistics and Data Analysis*, **50**, 830--858.
- Kenward, M.G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, **12**, 236--247.
- Kenward, M.G., Molenberghs, G., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete categorical data. *Statistical Modelling*, **1**, 31--48.

- Kenward, M.G., Molenberghs, G., and Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika*, **90**, 53–71.
- Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570–582.
- Li, K.H., Raghunathan, T.E., and Rubin, D.B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an *F* reference distributions. *Journal of the American Statistical Association*, **86**, 1065–1073.
- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
- Liang, K.-Y., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.
- Lipsitz, S.R., Laird, N.M., and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**, 153–160.
- Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.
- Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.
- Little, R.J.A., D'Agostino, R., Dickersin, K., Emerson, S.S., Farrar, J.T., Frangakis, C., Hogan, J.W., Molenberghs, G., Murphy, S.A., Neaton, J.D., Rotnitzky, A., Scharfstein, D., Shih, W., Siegel, J.P., and Stern, H. National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials*. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.
- Little, R.J.A. and Kang, S. (2015). Intention-to-treat analysis with treatment discontinuation and missing data in clinical trials. *Statistics in Medicine*, **34**, 2381–2390.
- Little, R.J.A. and Rubin, D.B. (2014). *Statistical Analysis with Missing Data* (3rd ed.). New York: John Wiley & Sons. [The first edition appeared in 1987; the second edition in 2002.]
- Lu, K. (2014). An analytic method for the placebo-based pattern-mixture model. *Statistics in Medicine*, **33**, 1134–1145.
- Mallinckrodt, C.H. (2013). *Preventing and Treating Missing Data in Longitudinal Clinical Trials: A Practical Guide*. New York: Cambridge University Press.
- Mallinckrodt, C.H., Clark, W.S., and Stacy R.D. (2001a). Type I error rates from mixed-effects model repeated measures versus fixed effects analysis of variance with missing values imputed via last observation carried forward. *Drug Information Journal*, **35**, 4, 1215–1225.
- Mallinckrodt, C.H., Clark, W.S., and Stacy R.D. (2001b). Accounting for dropout bias using mixed-effects models. *Journal of Biopharmaceutical Statistics*, **11**, (1 & 2), 9–21.
- Mallinckrodt, C.H. and Lipkovich, I. (2016). *Analyzing Longitudinal Clinical Trial Data. A Practical Guide*. Boca Raton: Chapman & Hall/CRC.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall.
- Michiels, B., Molenberghs, G., Bijnens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyze longitudinal quality of life data subject to dropout. *Statistics in Medicine*, **21**, 1023–1041.

- Michiels, B., Molenberghs, G., and Lipsitz, S.R. (1999). Selection models and pattern-mixture models for incomplete categorical data with covariates. *Biometrics*, **55**, 978–983.
- Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M.G. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B*, **70**, 371–388.
- Molenberghs, G., Fitzmaurice, G., Kenward, M.G., Verbeke, G., and Tsiatis, A.A. (2015). *Handbook of Missing Data*. Boca Raton: Chapman & Hall/CRC.
- Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.
- Molenberghs, G., Kenward, M.G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom dropout. *Biometrika* **84**, 33–44.
- Molenberghs, G., Kenward, M.G., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, **50**, 15–29.
- Molenberghs, G., Kenward, M.G., Verbeke, G., and Teshome Ayele, B. (2011). Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, **21**, 187–206.
- Molenberghs, G., Michiels, B., Kenward, M.G., and Diggle, P.J. (1998). Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, **52**, 153–161.
- Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M.G., Mallinckrodt, C., and Carroll, R.J. (2003). Analyzing incomplete longitudinal clinical trial data. *Submitted for publication*.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.
- Molenberghs, G., Verbeke, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001). Mastitis in dairy cattle: influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93–113.
- Murray G.D. and Findlay J.G. (1988). Correcting for the bias caused by drop-outs in hypertension trials. *Statistics in Medicine*, **7**, 941–946.
- O’Kelly, M. and Ratitch, B. (2014). *Clinical Trials with Missing Data: A Guide for Practitioners*. New York: John Wiley & Sons.
- Pharmacological Therapy for Macular Degeneration Study Group (1997). Interferon  $\alpha$ -IIA is ineffective for patients with choroidal neovascularization secondary to age-related macular degeneration. Results of a prospective randomized placebo-controlled clinical trial. *Archives of Ophthalmology*, **115**, 865–872.
- Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.
- Pinheiro, J. and Bates, D. M. (2000). *Mixed-effects Models in S and S-plus*. New York: Springer-Verlag.
- Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033–1048.
- Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C: the art of scientific computing*, Chapter 10. Cambridge University Press, New York, second edition.
- Robins, J.M., Rotnitzky, A., and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with non-ignorable non-response. *Journal of the American Statistical Association*, **93**, 1321–1339.

- Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.
- Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score method in observational studies for causal effects. *Biometrika*, **70**, 41–55.
- Rotnitzky, A., Cox, D.R., Bottai, M., and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, **6**, 243–284.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.
- Rubin, D.B. (1978). Multiple imputations in sample surveys -- a phenomenological Bayesian approach to nonresponse. In: *Imputation and Editing of Faulty or Missing Survey Data*. Washington, DC: U.S. Department of Commerce, pp. 1–23.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
- Rubin, D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, **81**, 366–374.
- Rubin, D.B., Stern H.S., and Vehovar V. (1995). Handling “don’t know” survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association*, **90**, 822–828.
- Schafer J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Scharfstein, D.O., Rotnitzky, A., and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, **94**, 1096–1120.
- Sheiner, L.B., Beal, S.L., and Dunne, A. (1997). Analysis of nonrandomly censored ordered categorical longitudinal data from analgesic trials. *Journal of the American Statistical Association*, **92**, 1235–1244.
- Skrondal, A. and Rabe-Hesketh, S. (2004). Generalized latent variable modeling: Multilevel, longitudinal and structural equation models. London: Chapman & Hall.
- Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550.
- Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., and Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics*, **3**, 245–265.
- Thijs, H., Molenberghs, G., and Verbeke, G. (2000). The milk protein trial: influence analysis of the dropout process. *Biometrical Journal*, **42**, 617–646.
- Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.
- Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D., Van den Noortgate, W., Meulders, M., and De Boeck, P. (2004). Estimation and software. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 343–373). New York: Springer.
- van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, **16**, 219–242.
- van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton: Chapman & Hall/CRC.
- van Buuren, S., Boshuizen, H.C., and Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, **18**, 681–694.

- Van Steen, K., Molenberghs, G., Verbeke, G., and Thijs, H. (2001). A local influence approach to sensitivity analysis of incomplete longitudinal ordinal data. *Statistical Modelling: An International Journal* **1**, 125--142.
- Verbeke, G., Lesaffre, E., and Brant L.J. (1998). The detection of residual serial correlation in linear mixed models. *Statistics in Medicine*, **17**, 1391--1402.
- Verbeke, G., Lesaffre, E., and Spiessens, B. (2001). The practical use of different strategies to handle dropout in longitudinal studies. *Drug Information Journal*, **35**, 419--434.
- Verbeke, G. and Molenberghs, G. (1997). *Linear Mixed Models in Practice: A SAS-Oriented Approach*. Lecture Notes in Statistics 126. New York: Springer-Verlag.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.
- Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001). Sensitivity analysis for non-random dropout: a local influence approach. *Biometrics*, **57**, 7--14.
- Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439--447.
- White, I.R., Daniel, R., and Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics and Data Analysis*, **54**, 2267--2275.
- Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233--243.
- Wu, M.C. and Carroll, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, **44**, 175--188.



# Index

## A

- Abelson, R.P. 135  
ADJUST statement 133  
adjusted analyses 2  
advanced randomization-based methods  
    about 67–68  
    analysis of binary endpoints 78–79  
    analysis of continuous endpoints using log-ratio  
        of two means 80–81  
    analysis of count endpoints using log-incidence  
        density ratios 81  
    analysis of ordinal endpoints using linear models  
        74–78  
    analysis of ordinal endpoints using proportional  
        odds model 79–80  
    analysis of time-to-event endpoints 82–86  
    case studies 70–72  
    nonparametric-based analysis of covariance 68–  
        70  
        %NParCov4 macro 73–74  
Agresti, A. 20, 28, 35, 128  
AIC (Akaike information criterion) 141  
Akaike information criterion (AIC) 141  
Allison, P.D. 4, 51  
Alosh, M. 221–222  
ALPHA option, %NParCov4 macro 95  
ALR (alternating logistic regressions) 332  
alternating logistic regressions (ALR) 332  
analysis  
    of binary endpoints 78–79  
    of continuous endpoints using log-ratio of two  
        means 80–81  
    of count endpoints using log-incidence density  
        ratios 81  
    of incomplete data 319–378  
    of ordinal endpoints using linear models 74–78  
    of ordinal endpoints using proportional odds  
        model 79–80  
    of time-to-event endpoints 82–86  
analysis of covariance, non-parametric  
    randomization-based 88–89  
ANCOVA models, Multiple Comparison-Modeling  
    (MCP-Mod) procedure based on 155–159  
Andersen, P.K. 42  
Anderson, G.L. 293, 297, 301  
Anderson, J.S. 293, 307  
ANOVA models, Multiple Comparison-Modeling  
    (MCP-Mod) procedure based on 145–155
- ANOVA procedure 6  
association, measures of 21  
asymptotic model-based tests 35–38  
asymptotic randomization-based tests 25–28
- B**
- Baker, S.G. 321, 363  
Bancroft, T.A. 15  
BASELINE statement 53  
Bates, D.M. 326  
Bauer, P.J. 252  
%BayesFutilityBin macro 310  
%BayesFutilityCont macro 305–307  
Bayesian CRM 110  
Bayesian framework 118  
Bayesian information criterion (BIC) 141  
Bayesian two-parameter logistic models,  
    implementation of 115–116  
Beal, S.L. 372  
Beckman, R.J. 366  
Benjamini, Y. 199  
%BetaModel macro 144  
Betensky, R.A. 293, 297, 301  
Beunckens, C. 338  
BIC (Bayesian information criterion) 141  
binary endpoints 78–79, 309–310  
Birch, M.W. 29  
Blackburn, P.R. 306–307  
Bonferroni method 184–187  
Bonferroni-based gatekeeping procedure 228–233  
Bortey, E.B. 15  
Brannath, W. 252, 313  
Branson, M. 116, 138, 140  
Brechenmacher, T. 238  
Breslow, N.E. 26–27, 29, 44, 54, 327  
Breslow-Day test 28, 61  
Bretz, F. 138, 140, 141, 148, 181, 209, 221–222, 237  
Buckland, S.T. 142  
BY statement 361  
Byar, D.P. 36
- C**
- C option, %NParCov4 macro 94  
calendar time 259  
candidate models, as a step in Multiple Comparison-  
Modeling (MCP-Mod) procedure 138–  
139  
Cantor, A. 4, 51  
Carey, V.C. 332

- Carpenter, J.R. 349, 353, 354, 363, 374–375  
 case studies  
   advanced randomization-based methods 70–72  
   categorical endpoints 21–22  
   continuous endpoints 4–6, 17  
   dose-finding methods 128–132, 144–176  
   incomplete data, analysis of 322–324  
   POCRM 118–123  
   repeated significance tests 257–258  
   stochastic curtailment tests 293–294  
 categorical endpoints  
   analysis of 20–40  
   asymptotic model-based tests 35–38  
   asymptotic randomization-based tests 25–28  
   case studies 21–22  
   exact model-based tests 38–40  
   exact randomization-based tests 28–32  
   minimum risk tests 32–34  
 CATMOD procedure, model-based inferences and 35  
 CC (complete case analysis) 320  
 %Chain macro 210  
 chain procedure 208–211, 244–249  
 Chakravorti, S.R. 15  
 Chang, C.K. 188, 201  
 Cheung, Y.K. 111, 114, 116, 119  
 Chinchilli, V.M. 15  
 Chuang-Stein, C. 128, 313  
 Ciminera, J.L. 15, 28, 57, 61  
 Claeskens, G. 141  
 CLASS statement 37–38, 52, 357  
 Clayton, D.G. 327  
 clinical information 182  
 closed family of hypotheses 189  
 closure principle 189–191  
 CMH (Cochran-Mantel-Haenszel) procedure 25–28  
 Cochran, W.G. 2, 25  
 Cochran-Armitage permutation test 29–30, 31  
 Cochran-Mantel-Haenszel (CMH) procedure 25–28, 29  
 Collett, D. 4, 44, 51  
 COMBINE option, %NParCov4 macro 94–95  
 complete case analysis (CC) 320  
 Conaway, M.R. 107, 117  
 conditional independence model 325  
 conditional power 294–312  
 continual reassessment method (CRM)  
   about 107–108  
   implementation of 111–116  
   modeling frameworks 108–111  
 continuous endpoints  
   analysis of 4–20, 80–81  
   analysis of using log-ratio of two means 80–81  
   case studies 17  
   fixed effects models 6–15  
   nonparametric tests 16–19  
   random effects models 15–16  
 CONTRAST statement 135–136  
 contrast-based tests 134–136  
 Cook, R.D. 365–367  
 copy reference 377  
 count endpoints, analysis of using log-incidence density ratios 81  
 covariance, nonparametric-based analysis of 68–70  
 COVARS option, %NParCov4 macro 94  
 Cox, D.R. 4, 51, 54  
 Cox proportional hazards model 2, 3  
 %CriticalValue macro 148–149, 158, 163, 166  
 CRM  
   *See* continual reassessment method (CRM)  
 cumulative cohort method 104  
 %Custom macro 238
- D**
- D'Agostino, R.B. 181, 206, 237  
 Daniel, R. 353  
 Danielson, L. 253  
 data monitoring committee charter 265  
 data-driven hypothesis ordering  
   about 189  
   closure principle 189–191  
   procedures with 189–202  
 Davis, B.R. 297, 303  
 Davis, C.S. 35  
 Day, N.E. 29, 36  
 DeCani, J.S. 261  
 decision matrix algorithm  
   about 191  
   Holm procedure 191–193  
   power comparisons 201–202  
   stepwise procedures 193–198  
   weighted procedures 199–201  
 DeMets, D.L. 253, 258, 264, 281, 289  
 Dempster, A.P. 337  
 DerSimonian, R. 321, 363  
 DETAILS option, %NParCov4 macro 96  
 Diggle P.J. 321, 332, 363–365  
 direct Bayesian analysis (ignorable Bayesian analysis) 341–344  
 DISCRETE method 54  
 distributional information 182, 183–184  
 DLT (dose-limiting toxicity) 102–103  
 Dmitrienko, A. 128, 137, 181–182, 184, 189, 191, 206, 221–222, 236–238, 289, 293–294, 298, 305, 309  
 dose-escalation methods  
   about 101  
   continual reassessment method (CRM) 107–116  
   partial order continual reassessment method 116–123  
   rule-based methods 103–107  
   trials 102–103  
 dose-finding methods  
   about 127–128

- case studies 128–132, 144–176
  - dose-response assessment and 132–144
- dose-limiting toxicity (DLT) 102–103
- dose-placebo tests 132–134
- dose-response analysis 127, 132–144
- double robustness 349
- %dropout macro 346
- %dropwgt macro 346
- DSNIN option, %NParCov4 macro 95
- DSNOUT option, %NParCov4 macro 95–96
- Dunne, A. 372
- Dunnett, C.W. 215
- Dunnett test 133–134, 140
- E**
  - Edwards, S. 29
  - efficiency robust tests 32
  - Efron, B. 54
  - Ellenberg, S.S. 253
  - %Emax macro 143, 146
  - Emerson, S.S. 275, 289
  - endpoints
    - See also* categorical endpoints
    - See also* continuous endpoints
    - binary 78–79, 309–310
    - count 81
    - normally distributed 305–306, 338–339
    - ordinal 74–80
    - primary 180
    - secondary 180
    - time-to-event 41–55, 43–51, 51–55, 82–86, 91–92
  - error spending functions 261–262
  - ESTIMATE statement 10, 336
  - estimation, following sequential testing 289–291
  - EXACT method 54
  - exact model-based tests 38–40
  - EXACT option, %NParCov4 macro 95
  - exact randomization-based tests 28–32
  - EXACT statement 23
  - EXACTONLY option, LOGISTIC procedure 38
  - exit probabilities 267
  - %Exponential macro 143, 146, 161–162
  - EXPOSURES option, %NParCov4 macro 94
- F**
  - F* statistic 14
  - fallback procedure 206–208
  - familywise error rate (FWER) 181–182
  - Faries, D. 111
  - Fisher's exact test for binary endpoints 2
  - Fitzmaurice, G.M. 320
  - fixed data monitoring strategies 258–259
  - fixed effects models 6–15
  - fixed-sequence procedure 203–206
  - Fleiss, J.L. 14–15, 25–26, 28
  - Fleming, T.R. 44, 253, 259, 275, 289
- flexible data monitoring strategies 258, 259–261
- Follman, D.A. 281
- frameworks
  - Bayesian 118
  - modeling 108–111
- Freedman, L.S. 306–307
- FREQ procedure
  - about 4
  - binary data and 29
  - obtaining test statistics using 27
  - odds ratio and 22–23
  - Output Delivery System (ODS) and 24–25
  - relative risk and 22–23
  - risk difference and 22–23
  - van Elteren statistics and 19
- Freund, J.L. 4, 7
- futility stopping rules 294
- FWDLINK statement 36
- FWER (familywise error rate) 181–182
- G**
  - Gail, M.H. 2, 28, 57–61
  - Gail-Simon test 57–61
  - Gallo, P.P. 14, 15, 253
  - Gastwirth, J.L. 32
  - gatekeeping procedures
    - about 221–222
    - implementation of mixture-based procedures 227–236
  - inferences based on mixture method 223–236
  - modified mixture method 236–240
  - ordered families of hypotheses in clinical trials 222–223
- Gatsonis, C. 116
- GEE
  - See* generalized estimating equations (GEE)
- GEE procedure
  - data analysis 320
  - incomplete data analysis 341, 348–349, 357, 360, 361
  - LOCF and 338
- Gehan, E.A. 44
- Geisser, S. 293
- generalized estimating equations (GEE)
  - about 320, 327, 331–332
  - methodological detail 332–333
  - weighted (WGEE) 344–349
- generalized linear mixed models (GLMM) 325–330
- %GeneralizedOptimalContrasts macro 147, 155–156, 162, 166
- GENMOD procedure
  - about 4
  - covariance matrix and 156
  - data analysis 320
  - dose-placebo tests 133
  - incomplete data analysis 341, 346–347, 357, 360, 361

- interim data monitoring 300
- LOCF and 338
- model-based inferences and 35, 37–38
- Genz, A. 141, 148
- GLIMMIX procedure
  - data analysis 320
  - generalized linear mixed models 329–330
  - incomplete data analysis 333, 341, 350, 356, 357, 360, 361
  - LOCF and 338
  - parametric procedures 215
  - step-down Dunnett procedure 218–220
- GLM procedure
  - about 4, 6, 8–9
  - Bonferroni method 184
  - dose-placebo tests 133
  - dose-response tests 148, 157
  - ESTIMATE statement 10
  - parametric procedures 215
  - Sidak method 184
  - in Type II, II, and III analysis 13–15
  - Type III analysis and 12
- GLMM (generalized linear mixed models) 325–330
- global null hypothesis 181
- Glynn, R.J. 321
- Goldberg, J.D. 4
- Goodman, S.N. 111
- Gould, A.L. 4
- Govindarajulu, Z. 289
- Greenhouse, J.B. 116
- Greenland, S. 26–27
- Grizzle, J.E. 15
- group sequential designs
  - about 251, 254
  - comparing 262–263
  - for detecting futility 255
  - for detecting superior efficacy 254–255
  - for simultaneous efficacy and futility testing 255–256
- Gsponder, T. 116
- Guo, W. 105
- H**
- Hackney, O.P. 6
- Haenszel, W. 25, 26
- Hájek, J. 44
- Halperin, M. 292, 295, 296
- Hardy, R.J. 297, 303
- Harrington, D.P. 44, 332
- Harrington-Fleming test 44–45, 49, 51
- Hartley, H.O. 15
- Hartzel, J. 20
- Herson, J. 307, 310
- Hives Severity (HS) score 129–132
- Hjorth, N.L. 141
- Hochberg, Y. 181, 199
- Hochberg procedure 193, 197–198
- Hocking, R.R. 6, 11
- Hogan, J.W. 371–372
- Holm, S. 199
- Holm procedure 191–193
- Hommel, G. 195
- Hommel procedure 193, 195–197, 245
- Hommel-based gatekeeping procedure 233–236, 246–249
- Hosmane, B. 17
- HS (Hives Severity) score 129–132
- Hsu, J.C. 181
- Hunsberger, S.A. 313
- Hwang, I.K. 261
- HYPOTH option, %NParCov4 macro 95
- hypothesis testing 352
- I**
- ignorability 331
- ignorable likelihood (direct likelihood) 338–341
- implication relationships 189–190
- imputation
  - mechanisms for 353–356
  - simple methods of 337–338
  - as a step 354
  - strategies for 166–176
- incomplete data, analysis of
  - case studies 322–324
  - data setting and methodology 324–333
  - direct Bayesian analysis (ignorable Bayesian analysis) 341–344
  - ignorable likelihood (direct likelihood) 338–341
  - multiple imputation 349–362
  - sensitivity analysis 362–378
  - sensitivity analysis based on multiple imputation and pattern-mixture models 371–378
  - sensitivity analysis using local influence 363–371
  - simple methods and MCAR 334–338
  - weighted generalized estimating equations 344–349
- Inference Task 351
- information time 259
- initial step 354
- interim data monitoring
  - about 251–253
  - repeated significance tests 253–292
  - stochastic curtailment tests 293–315
- INVAR procedure 34
- INVAR test 32, 33
- inverse probability weighting (IPW) 349
- IPW (inverse probability weighting) 349
- Ivanova, A. 104
- J**
- Jansen, I. 321, 363
- Jennison, C. 252, 254, 260, 261, 286–287, 290, 297
- Ji, Y. 104–107

Johns, D. 293, 307  
 Johnson, D.E. 4, 6, 7

Johnson, W. 293  
 Jones, B. 14, 15

## K

Kalbfleisch, J.D. 4, 48, 51  
 Kaplan-Meier estimate 42–43  
 Kass, R.E. 142  
 Kawaguchi, A. 68  
 Kenward, M.G. 321, 334, 349, 353–354, 363–365,  
     367, 371–372, 374–375  
 Kim, K. 289  
 Klein, J.P. 42  
 Koch, G.G. 17, 29, 35, 68–69  
 Koury, K.J. 4

## L

Lachin, J.M. 2, 4, 20, 35, 44  
 Lagakos, S.W. 2  
 Laird, N.M. 321, 332, 371–372  
 Lan, K.K.G. 252, 253, 258, 264, 281, 292, 295–296  
 Laplace method 329  
 large-strata asymptotics 28–29  
 last observation carried forward (LOCF) 320, 337–  
     338  
 LaVange, L. 253  
 Lee, J.J. 293  
 Lee, S.M. 111, 119  
 Lehmann, E.L. 17  
 Lesaffre, E. 321, 366  
 Li, K.H. 352  
 Liang, K.-Y. 331, 333, 344  
 LIFEREG procedure 51  
 LIFETEST procedure  
     about 4  
     inferences performed by 50–51  
     Kaplan-Meier estimate and 42–43  
     qualitative information from 53  
     STRATA statement 45–46  
     TEST statement 46–48  
     warning messages and 39  
 likelihood ratio test, randomization-based tests and  
     43–51  
 likelihood-based approaches 330–331  
 linear mixed models 324–325  
 linear models, analysis of ordinal endpoints using  
     74–78  
 linear rank tests 44  
 %LinRank macro 49, 51  
 Lipkovich, I. 166, 350, 355  
 Lipsitz, S.R. 332  
 Littell, R.C. 4, 7, 15  
 Little, R.J.A. 320, 324, 331, 337, 349, 371  
 Liu, D.D. 293  
 Liu, W. 189

LOCF (last observation carried forward) 320  
 Logan, B.R. 137  
 logical restrictions 182, 183  
 log-incidence density ratios 81, 90–91  
 %Logistic macro 143–144  
 LOGISTIC procedure  
     about 4  
     exact inferences and 38–40  
     model-based inferences and 35, 37–38  
 logistic regression models 3  
 logistic transformation 89–90  
 logit-adjusted estimate 26  
 LogLinear model 149–156  
 log-rank scores, for time-to-event outcomes 91–92  
 log-rank test  
     randomization-based tests and 43–51  
     for time-to-event end points 2  
 log-ratio of two means, analysis of continuous  
     endpoints using 80–81  
 log-ratio of two outcome vectors 90  
 Lu, K. 375

## M

Mallinckrodt, C.H. 166, 331, 350, 355  
 Mantel, N. 25, 26, 29  
 Mantel-Fleiss criterion 29  
 Mantel-Haenszel test 3, 17, 26, 28, 44  
 MAR (missing at random) 320  
 marginal quasi-likelihood estimates (MQL) 327  
 MCAR (missing completely at random) 320  
 MCAR, simple methods and 334–338  
 MCMC procedure 350, 354, 355  
 MCP-Mod  
     *See* Multiple Comparison-Modeling (MCP-Mod)  
     procedure  
 MED (minimum effective dose) 128, 141–142, 152–  
     156  
 median unbiased estimate 290  
 Mehrotra, D.V. 28, 32, 33  
 Mehta, C.R. 289–290  
 Menon, S. 128  
 MI  
     *See* multiple imputation (MI)  
 MI procedure  
     incomplete data analysis 356–359  
     NIMPUTE option 168–169  
 MIANALYZE procedure 169, 350, 356, 358, 359,  
     362  
 Milliken, G.A. 4, 6, 7, 15  
 minimum effective dose (MED) 128, 141–142, 152–  
     156  
 "minimum regret" procedure 33  
 minimum risk tests 32–34  
 %MinRisk macro 33  
 missing at random (MAR) 320  
 missing completely at random (MCAR) 320  
 missing not at random (MNAR) 168

- missing observations, Multiple Comparison-Modeling (MCP-Mod) procedure in presence of 165–176
- MISSMODEL statement 348
- MIXED procedure
- about 4, 6
  - analysis of ordinal endpoints using linear models 76
  - Bonferroni method 184
  - contrast-based tests 135
  - data analysis 320
  - dose-placebo tests 133
  - dose-response tests 163, 172
  - incomplete data analysis 341, 350, 357, 360
  - LOCF and 338
  - parametric procedures 215
  - random effects models and 15–16
  - Šidák method 184
  - in Type II, II, and III analysis 13–15
  - Type III analysis and 12
- %MixGate macro 227–229, 236–238
- mixture method, general principles of gatekeeping inferences based on 223–226
- mixture-based gatekeeping procedures, implementation of 227–236
- MNAR (missing not at random) 168
- MNAR statement 376
- model selection, as a step in Multiple Comparison-Modeling (MCP-Mod) procedure 141
- MODEL statement 53, 325, 330
- model-based tests 1–4, 51–55
- modeling frameworks 108–111
- modified mixture method 236–240, 247–249
- modified toxicity probability interval method (mTPI) 104–107
- Molenberghs, G. 320–321, 325, 332, 334, 344, 349, 363, 367–368, 371–372
- MONOTONE statement 353, 354, 357
- Morris, D. 17
- MQL (marginal quasi-likelihood estimates) 327
- mTPI (modified toxicity probability interval method) 104–107
- MTPs (multiple testing procedures) 180–184
- multi-contrast tests 136–137
- Multiple Comparison-Modeling (MCP-Mod)
- procedure
  - about 128, 137–138
  - based on ANCOVA models 155–159
  - based on repeated-measures models 161–165
  - based on simple ANOVA models 145–155
  - candidate models 138–139
  - dose-finding on 138
  - dose-response tests 140–141
  - implementation of 143–144
  - model selection 141
  - optimal contrasts 139–140
  - in presence of missing observations 165–176
- target dose selection 141–142
- multiple imputation (MI)
- about 349–350
  - efficiency 352–353
  - hypothesis testing 352
  - imputation mechanisms 353–356
  - pooling information 351–352
  - SAS code for 358–362
  - SAS for 356–358
  - sensitivity analysis based on pattern-mixture models and 371–378
  - theoretical justification 350–351
- multiple testing procedures (MTPs) 180–184
- multiplicity adjustment methods
- about 179–181
  - control of familywise error rate 181–182
  - gatekeeping procedures 221–241
  - multiple testing procedures 182–184
  - parametric procedures 212–221
  - procedures with data-driven hypothesis ordering 189–202
  - procedures with prespecified hypothesis ordering 202–212
  - single-step procedures 184–188
- multivariate modeling 354
- %MultSeqSimul macro 237
- MULTTEST procedure
- Bonferroni method 184, 193
  - Cochran-Armitage permutation and 39
  - Fisher test and 29–30
  - Hochberg procedure 197–198
  - Holm procedure 193
  - Hommel procedure 196
  - permutations and 31–32
  - resampling-based exact methods 93
  - Šidák method 184
  - Simes method 187
  - weighted procedures 200
- N**
- Nachtsheim, C.J. 366
- Nelder, J.A. 14
- Neuenschwander, B. 116
- NIMPUTE option, MI procedure 168–169
- NLMIXED procedure
- data analysis 320
  - generalized linear mixed models 325, 329–330
  - incomplete data analysis 341, 350, 356, 357, 360, 361
  - LOCF and 338
  - model selection 149, 158–159, 164, 173–174
  - model-based inferences and 35
  - target dose selection 165
- non-Gaussian outcomes 339–340
- non-parametric randomization-based analysis of covariance 88–89
- nonparametric tests 16–19

- nonparametric-based analysis of covariance 68–70
  - non-prognostic factors 1–2
  - non-random (MNAR) 320
  - non-responder imputation 166–167
  - non-testable hypotheses 224
  - normally distributed endpoints/outcomes 305–306, 338–339
  - %NParCov4 macro
    - about 73–74
    - content of output data sets 96–99
    - general use and options for 94–96
  - NREPS option, %NParCov4 macro 95
- O**
- Oakes, D.O. 4, 51
  - O'Brien, P.C. 259, 275
  - O'Brien-Fleming stopping boundary 258, 263–270, 275–276, 280–284, 287–288
  - O'Connell, M. 327
  - odds ratio 21
  - ODS (Output Delivery System), FREQ procedure
    - and 24–25
  - ODS statement 358, 361
  - Öhashi, Y. 42
  - O'Kelly, M. 166, 167, 349
  - one-parameter logistic model 108–110
  - optimal contrasts, as a step in Multiple Comparison Modeling (MCP-Mod) procedure 139–140
  - O'Quigley, J. 107, 117
  - ordered families of hypotheses in clinical trials 222–223
  - ordinal endpoints
    - analysis of using linear models 74–78
    - analysis of using proportional odds model 79–80
  - OUTCOMES option, %NParCov4 macro 94
  - Output Delivery System (ODS), FREQ procedure
    - and 24–25
- P**
- Pampallona, S. 261
  - PANSS (Positive and Negative Syndrome Scale) 128–129
  - parametric procedures
    - about 212–214
    - single-step Dunnett procedure 214–217
    - stepwise Dunnett procedures 217–220
  - partial likelihood 51
  - partial order 116–117
  - partial order continual reassessment method (POCRM)
    - about 107–108, 116–117
    - Bayesian framework 118
    - in case studies 118–123
    - partial order 117
  - partitioning principle 191
  - patient populations, multiple 180
  - patient's latent toxicity 119
  - Patra, K. 313
  - pattern-mixture imputation, with placebo imputation 167–176
  - pattern-mixture models (PMM) 320, 371–378
  - penalized quasi-likelihood estimates (PQL) 327
  - Pepe, M.S. 293, 297, 301
  - permutation *t*-test for continuous endpoints 2
  - Peto, J. 44
  - Peto, R. 44
  - PHREG procedure
    - about 4, 41–42
    - proportional hazards models and 51–53
    - ties and 54
  - Piantadosi, S. 2, 111
  - Pinheiro, J.C. 138, 140, 155, 326
  - PMM (pattern-mixture models) 320, 371–378
  - PMM (predictive mean matching) 353
  - Pocock, S.J. 259
  - Pocock stopping boundary 258, 271–276, 280–281, 284–285
  - POCRM
    - See* partial order continual reassessment method (POCRM)
  - %POCRM macro 120–121
  - Poisson regression methods 3
  - pooling information 351–352
  - Positive and Negative Syndrome Scale (PANSS) 128–129
  - posterior step 354
  - power comparisons 201–202
  - power model 108–110
  - PQL (penalized quasi-likelihood estimates) 327
  - precision 2
  - predictive mean matching (PMM) 353
  - predictive power 294–312
  - predictive probability, futility rules based on 304–309
  - Prentice, R.L. 4, 44, 48, 51, 332
  - prespecified hypothesis ordering
    - about 202–203
    - chain procedure 208–211
    - fallback procedure 206–208
    - fixed-sequence procedure 203–206
    - procedures with 202–212
  - pre-testing, Type III analysis with 14–15
  - primary endpoints, multiple 180
  - PROBIT procedure, model-based inferences and 35
  - prognostic factors 1
  - propensity scores 353
  - proportional odds model 3, 79–80
  - Proschan, M.A. 252, 281, 313
  - pseudo-quasi-likelihood 327
  - Pulkstenis, E. 236, 238, 294
  - pushback test 61
  - %PvalProc macro 200, 207
  - p*-values 225–226

**Q**

Qaqish, B. 332  
%Quadratic macro 144  
qualitative interaction tests  
  about 56–57, 61  
  Gail-Simon test 57–61  
quantitative interactions 15

**R**

*R()* notation  
  Type I analysis and 7–9  
  Type II analysis and 9–11  
Rabe-Hesketh, S. 326  
Radhakrishna, S. 32, 33  
Raftery, A.E. 142  
Raghunathan, T.E. 352  
Railkar, R. 32, 33  
random effects models 15–16  
RANDOM statement 320, 330, 341  
randomization-based methods 1–4  
  *See also* advanced randomization-based methods  
randomization-based tests 43–51  
Rao J.N.K. 4, 15  
Ratitch, B. 166, 167, 349  
%RegularOptimalContrasts macro 146–147  
relative risk 21  
REML (restricted maximum likelihood) estimation  
  327  
repeated confidence intervals 286–288  
repeated significance tests  
  about 253  
  case studies 257–258  
  design and monitoring stages 258  
  design stage 263–280  
  estimation following sequential testing 289–291  
  fixed and flexible data monitoring strategies  
  258–263  
  group sequential trial designs 254–256  
  monitoring stage 280–285  
  relationship with conditional power tests 297–  
  298  
  repeated confidence intervals 286–288  
REPEATED statement 325, 330, 333, 347  
repeated-measures models, Multiple Comparison-  
  Modeling (MCP-Mod) procedure based on  
  161–165  
resampling-based exact methods 93  
responder imputation 167  
restricted maximum likelihood (REML) estimation  
  327  
risk difference 21  
Robins, J.M. 26–27, 321, 344, 363  
Rodriguez, R. 14  
Roger, J.H. 363, 374  
Rosenberger, W.F. 321, 363  
Rosner, G.L. 289–290  
Rotnitzky, A. 321, 344, 363

Royston, P. 353

Ruberg, S.J. 128, 137  
Rubin, D.B. 320–321, 324, 331, 337, 349, 352, 355,  
  361  
Rüger, B. 187  
rule-based methods  
  about 103  
  modified toxicity probability interval method  
    (mTPI) 104–107  
  up-and-down methods 103–104

**S**

Sarkar, S.K. 188, 201  
Saville, B.R. 68  
Schafer, J.L. 349, 354  
Scharfstein, D.O. 321, 363  
Scheffe, H. 6, 135  
Schenker, N. 349  
Schoenfeld, D.A. 2  
Searle, S.R. 4, 7, 15  
secondary endpoints, multiple 180  
SEED option, %NParCov4 macro 95  
selection modeling 320  
Senn, S. 15, 16, 57, 237  
sensitivity analysis  
  about 362–363  
  based on multiple imputation and pattern-mixture  
    models 371–378  
  using local influence 363–371  
separable procedures 225  
SEQDESIGN procedure, interim data monitoring  
  252–253, 258, 261, 264, 271, 277, 299  
%SeqPower macro 302–303  
SEQTEST procedure, interim data monitoring 252–  
  253, 258, 261, 271, 281–285, 287, 290–  
  292, 297, 298, 300–301  
sequential rejection principle 191  
sequential testing, estimation following 289–291  
Sheiner, L.B. 372  
Shih, W.J. 261  
Shu, V. 17  
Šidák, Z. 44  
Šidák method 184–187  
Siegmund, D. 289–290  
%SigEmax macro 144, 161–162  
Simes method 187–188  
Simon, R. 28, 57–61, 292, 295–296  
simple order 116  
single-contrast tests 136  
single-step Dunnett procedure 214–217  
single-step procedures  
  about 184  
  Bonferroni method 184–187  
  Šidák method 184–187  
  Simes method 187–188  
Skrondal, A. 326  
sparse-data asymptotics 28–29

- Spector, P.C. 4, 7  
 Speed, F.M. 6, 11  
 Spiegelhalter, D.J. 306–307  
 SQL procedure, dose-response tests 147–148, 157  
 SSIZE procedure 33, 34  
 standard mixture method, Hommel-based  
     gatekeeping procedure using 246–247  
 %StartUp macro 144, 145, 162  
 statistical information 182  
 STDERR statement 358  
 step-down algorithm 193–195  
 step-down Dunnett procedure 217–220  
 step-up Dunnett procedure 220  
 stepwise Dunnett procedures 217–220  
 stepwise procedures 193–198  
 Stern, H.S. 321  
 Stewart, W.H. 137  
 stochastic curtailment tests  
     about 292–293  
     applications of 312–313  
     case studies 293–294  
     futility rules based on conditional and predictive power 294–312  
 Stokes, M.E. 35  
 stopping probabilities 267  
 Storer, B.E. 115–116  
 STRATA option, %NParCov4 macro 94  
 STRATA statement 39, 45–46, 52  
 stratification 92  
 stratified analysis, of time-to-event data 50–51  
 stratified log-rank test 3  
 strong control 182  
 Stroup, W.W. 15  
 SYMSIZE option, %NParCov4 macro 96
- T**
- TABLE statement 19  
 Tamhane, A.C. 137, 138, 181, 222  
 Tan, W.Y. 2  
 Tangen, C.M. 68, 69  
 Tanner, M.A. 349  
 target dose selection, as a step in Multiple Comparison-Modeling (MCP-Mod) procedure 141–142  
 Tarone, R.E. 44, 45, 49  
 Tarone-Ware test 44–45, 49, 51  
 TEST statement, LIFETEST procedure 46–48, 358  
 testable hypotheses 224  
 Thijss, H. 321, 372  
 ties, analysis of time-to-event data with 53–55  
 time-to-event data  
     analysis of with ties 53–55  
     stratified analysis of 50–51  
 time-to-event endpoints/outcomes  
     analysis of 41–55, 82–86  
     log-rank scores for 91–92  
     model-based tests 51–55
- randomization-based tests 43–51  
 Wilcoxon scores for 91–92  
 Ting, N. 128  
 tipping point analysis 375  
 Tobias, R.D. 14  
 TRANSFORM option, %NParCov4 macro 95  
 TRANSFORM statement 357  
 treatment comparisons, multiple 179–180  
 TRTGRPS option, %NParCov4 macro 94  
 truncated Hommel procedure 245  
 truncated procedures 225  
 Tsiatis, A.A. 261, 275, 289–290  
*t*-statistic method 104  
 Tuerlinckx, F. 326  
 Tukey, J.W. 135, 137  
 Turnbull, B.W. 252, 254, 260–261, 286–287, 290, 297  
 Type I analysis 7–9, 13–15  
 Type II analysis 9–11, 13–15  
 Type II error rate control 303  
 Type III analysis  
     about 11–13  
     compared with Type I and Type III 13–15  
     with pre-testing 14–15
- U**
- unit probability mass (UPM) 105–107  
 up-and-down methods 103–104  
 UPM (unit probability mass) 105–107
- V**
- van Buuren, S. 349  
 Van Den Berghe, G. 289  
 van Elteren, P.H. 16–17  
 Van Elteren test 3  
 Van Steen, K. 321  
 Vansteelandt, S. 349  
 Vehovar, V. 321  
 Verbeke, G. 321, 325, 332, 334, 366
- W**
- Waclawiw, M.A. 281  
 Wages, N.A. 107, 117  
 Wald chi-square statistic 36  
 Wang, M.D. 293, 298, 309, 313  
 Wang, S.J. 105  
 Wang, S.K. 275  
 Wang, Y. 313  
 Ware, J. 44–45, 49  
 Wassmer, G. 252, 313  
 Wedderburn, R.W.M. 331  
 weighted generalized estimating equations (WGEE) 344–349  
 weighted procedures 199–201  
 Westfall, P.H. 29, 181, 184, 192  
 White, I.R. 353  
 Whitehead, J. 253, 289

- Wieand, S. 2  
Wilcoxon rank sum test 2, 16–17, 43–51  
Wilcoxon scores, for time-to-event outcomes 91–92  
Wittes, J. 252, 297  
Wolfinger, R.D. 14, 15, 327  
Wong, W.H. 349  
working correlation matrix 332

**Y**

- Yamaguchi, T. 42  
Young, S.S. 184, 192

**Z**

- Zeger, S.L. 331–333, 344  
Zhang, M. 42  
Zhao, L.P. 344  
Zink, R. 128

# Ready to take your SAS® and JMP® skills up a notch?



Be among the first to know about new books,  
special events, and exclusive discounts.

[support.sas.com/newbooks](http://support.sas.com/newbooks)

Share your expertise. Write a book with SAS.

[support.sas.com/publish](http://support.sas.com/publish)

 [sas.com/books](http://sas.com/books)  
for additional books and resources.

  
THE POWER TO KNOW®

