

# **Multiple Testing Problems in Pharmaceutical Statistics**

# Chapman & Hall/CRC Biostatistics Series

Editor-in-Chief

**Shein-Chung Chow, Ph.D.**

*Professor*

*Department of Biostatistics and Bioinformatics*

*Duke University School of Medicine*

*Durham, North Carolina, U.S.A.*

Series Editors

**Byron Jones**

*Senior Director*

*Statistical Research and Consulting Centre  
(IPC 193)*

*Pfizer Global Research and Development  
Sandwich, Kent, U. K.*

**Jen-pei Liu**

*Professor*

*Division of Biometry  
Department of Agronomy  
National Taiwan University  
Taipei, Taiwan*

**Karl E. Peace**

*Georgia Cancer Coalition*

*Distinguished Cancer Scholar*

*Senior Research Scientist and*

*Professor of Biostatistics*

*Jiann-Ping Hsu College of Public Health  
Georgia Southern University  
Statesboro, Georgia*

**Bruce W. Turnbull**

*Professor*

*School of Operations Research  
and Industrial Engineering  
Cornell University  
Ithaca, New York*

# Chapman & Hall/CRC Biostatistics Series

## Published Titles

1. *Design and Analysis of Animal Studies in Pharmaceutical Development*, Shein-Chung Chow and Jen-pei Liu
2. *Basic Statistics and Pharmaceutical Statistical Applications*, James E. De Muth
3. *Design and Analysis of Bioavailability and Bioequivalence Studies, Second Edition, Revised and Expanded*, Shein-Chung Chow and Jen-pei Liu
4. *Meta-Analysis in Medicine and Health Policy*, Dalene K. Stangl and Donald A. Berry
5. *Generalized Linear Models: A Bayesian Perspective*, Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick
6. *Difference Equations with Public Health Applications*, Lemuel A. Moyé and Asha Seth Kapadia
7. *Medical Biostatistics*, Abhaya Indrayan and Sanjeev B. Sarmukaddam
8. *Statistical Methods for Clinical Trials*, Mark X. Norleans
9. *Causal Analysis in Biomedicine and Epidemiology: Based on Minimal Sufficient Causation*, Mikel Aickin
10. *Statistics in Drug Research: Methodologies and Recent Developments*, Shein-Chung Chow and Jun Shao
11. *Sample Size Calculations in Clinical Research*, Shein-Chung Chow, Jun Shao, and Hansheng Wang
12. *Applied Statistical Design for the Researcher*, Daryl S. Paulson
13. *Advances in Clinical Trial Biostatistics*, Nancy L. Geller
14. *Statistics in the Pharmaceutical Industry, Third Edition*, Ralph Buncher and Jia-Yeong Tsay
15. *DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments*, David B. Allison, Grier P. Page, T. Mark Beasley, and Jode W. Edwards
16. *Basic Statistics and Pharmaceutical Statistical Applications, Second Edition*, James E. De Muth
17. *Adaptive Design Methods in Clinical Trials*, Shein-Chung Chow and Mark Chang
18. *Handbook of Regression and Modeling: Applications for the Clinical and Pharmaceutical Industries*, Daryl S. Paulson
19. *Statistical Design and Analysis of Stability Studies*, Shein-Chung Chow
20. *Sample Size Calculations in Clinical Research, Second Edition*, Shein-Chung Chow, Jun Shao, and Hansheng Wang
21. *Elementary Bayesian Biostatistics*, Lemuel A. Moyé
22. *Adaptive Design Theory and Implementation Using SAS and R*, Mark Chang
23. *Computational Pharmacokinetics*, Anders Källén
24. *Computational Methods in Biomedical Research*, Ravindra Khattree and Dayanand N. Naik
25. *Medical Biostatistics, Second Edition*, A. Indrayan
26. *DNA Methylation Microarrays: Experimental Design and Statistical Analysis*, Sun-Chong Wang and Arturas Petronis
27. *Design and Analysis of Bioavailability and Bioequivalence Studies, Third Edition*, Shein-Chung Chow and Jen-pei Liu
28. *Translational Medicine: Strategies and Statistical Methods*, Dennis Cosmatos and Shein-Chung Chow
29. *Bayesian Methods for Measures of Agreement*, Lyle D. Broemeling
30. *Data and Safety Monitoring Committees in Clinical Trials*, Jay Herson
31. *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*, Karl E. Peace
32. *Bayesian Missing Data Problems: EM, Data Augmentation and Noniterative Computation*, Ming T. Tan, Guo-Liang Tian, and Kai Wang Ng
33. *Multiple Testing Problems in Pharmaceutical Statistics*, Alex Dmitrienko, Ajit C. Tamhane, and Frank Bretz

Chapman & Hall/CRC Biostatistics Series

# Multiple Testing Problems in Pharmaceutical Statistics

Edited by

**Alex Dmitrienko**

Eli Lilly and Company  
Indianapolis, Indiana, U.S.A.

**Ajit C. Tamhane**

Northwestern University  
Evanston, Illinois, U.S.A.

**Frank Bretz**

Novartis Pharma AG  
Basel, Switzerland

Hannover Medical School  
Germany



**CRC Press**

Taylor & Francis Group  
Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

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

© 2010 by Taylor and Francis Group, LLC  
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper  
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-58488-984-7 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

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

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

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

---

#### Library of Congress Cataloging-in-Publication Data

---

Multiple testing problems in pharmaceutical statistics / editors, Alex Dmitrienko, Ajit C. Tamhane, Frank Bretz.  
p. ; cm. -- (Chapman & Hall/CRC biostatistics series ; 33)  
Includes bibliographical references and index.  
ISBN 978-1-58488-984-7 (hardcover : alk. paper)  
1. Drugs--Testing--Statistical methods. I. Dmitrienko, Alex. II. Tamhane, Ajit C. III. Bretz, Frank. IV. Series: Chapman & Hall/CRC biostatistics series ; 33.  
[DNLM: 1. Clinical Trials as Topic. 2. Statistics as Topic. QV 771 M957 2010]

RM301.27.M85 2010  
615'.1901--dc22

2009040277

---

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

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

---

# Contents

<b>Preface</b>	<b>xi</b>
<b>List of Contributors</b>	<b>xv</b>
<b>1 Multiplicity Problems in Clinical Trials: A Regulatory Perspective</b>	<b>1</b>
<i>Mohammad Huque and Joachim Röhmel</i>	
1.1 Introduction	1
1.2 Common multiplicity problems in clinical trials	3
1.2.1 Clinical win criteria with primary endpoints	4
1.2.2 Consistency-ensured analysis of multiple endpoints	7
1.2.3 Co-primary endpoints	8
1.2.4 Secondary endpoints	11
1.2.5 Dose-response and dose-control comparisons	12
1.2.6 Subgroup analyses	14
1.2.7 Combination drug trials	16
1.2.8 Three-arm trials with an active control	17
1.2.9 Advanced multiplicity problems	20
1.3 Reducing multiplicity in clinical trials	23
1.3.1 Hierarchical testing	23
1.3.2 Composite endpoints	25
1.4 Multiplicity concerns in special situations	28
1.4.1 Use of multiple subsets of patient data	28
1.4.2 Use of multiple statistical methods	29
1.4.3 Analysis of covariance	29
1.5 Multiplicity in the analysis of safety endpoints	30
1.6 Concluding remarks	32
<b>2 Multiple Testing Methodology</b>	<b>35</b>
<i>Alex Dmitrienko, Frank Bretz, Peter H. Westfall, James Troendle, Brian L. Wiens, Ajit C. Tamhane, and Jason C. Hsu</i>	
2.1 Introduction	35
2.2 Error rate definitions	36
2.2.1 Comparisonwise and familywise error rates	36
2.2.2 Generalized familywise error rate, false discovery rate and false discovery proportion	38

2.2.3	Role of Type II errors . . . . .	39
2.3	Multiple testing principles . . . . .	40
2.3.1	Union-intersection testing . . . . .	40
2.3.2	Intersection-union testing . . . . .	41
2.3.3	Closure principle . . . . .	41
2.3.4	Properties of closed testing procedures . . . . .	44
2.3.5	Partitioning principle . . . . .	45
2.4	Adjusted significance levels, $p$ -values and confidence intervals . . . . .	46
2.4.1	Adjusted significance levels and $p$ -values . . . . .	46
2.4.2	Simultaneous confidence intervals . . . . .	47
2.5	Common multiple testing procedures . . . . .	48
2.5.1	Classification of multiple testing procedures . . . . .	48
2.5.2	Notation . . . . .	50
2.5.3	Dose-finding trial example . . . . .	50
2.6	Multiple testing procedures based on univariate $p$ -values . . . . .	52
2.6.1	Bonferroni procedure . . . . .	52
2.6.2	Holm procedure and its extensions . . . . .	53
2.6.3	Fixed-sequence procedure . . . . .	56
2.6.4	Fallback procedure . . . . .	57
2.6.5	Bonferroni-based closed testing procedures . . . . .	58
2.6.6	Reverse fixed-sequence procedure . . . . .	62
2.6.7	Simes global test . . . . .	63
2.6.8	Hommel procedure . . . . .	65
2.6.9	Hochberg procedure . . . . .	66
2.6.10	Adjusted $p$ -values . . . . .	67
2.6.11	Simultaneous confidence intervals . . . . .	70
2.7	Parametric multiple testing procedures . . . . .	73
2.7.1	Single-step Dunnett procedure . . . . .	74
2.7.2	Stepwise Dunnett procedures . . . . .	75
2.7.3	Extended Shaffer-Royen procedure . . . . .	78
2.7.4	Adjusted $p$ -values and simultaneous confidence intervals . . . . .	78
2.7.5	Multiple comparisons in general linear models . . . . .	81
2.8	Resampling-based multiple testing procedures . . . . .	82
2.8.1	Closed resampling-based procedures . . . . .	83
2.8.2	Step-down Dunnett procedures based on parametric and nonparametric resampling . . . . .	85
2.8.3	Step-down resampling-based procedures for multivariate linear models . . . . .	88
2.8.4	Permutation-based procedures . . . . .	90
2.9	Software implementation . . . . .	94
2.9.1	Multiple comparison procedures in SAS . . . . .	94
2.9.2	Multiple comparison procedures in R . . . . .	95

<b>3</b>	<b>Multiple Testing in Dose-Response Problems</b>	<b>99</b>
	<i>Frank Bretz, Ajit C. Tamhane, and José Pinheiro</i>	
3.1	Introduction . . . . .	99
3.2	Dose-response trend tests . . . . .	101
3.2.1	Notation . . . . .	101
3.2.2	Contrast tests . . . . .	102
3.2.3	Multiple contrast tests for general linear models . . . .	104
3.2.4	Thalidomide study example . . . . .	105
3.3	Target dose estimation using multiple hypothesis testing . .	109
3.3.1	Estimation of the minimum effective dose . . . . .	111
3.3.2	Estimation of the maximum safe dose . . . . .	115
3.4	Power and sample size calculation for target dose estimation	117
3.5	Hybrid approaches combining multiple testing and modeling	118
3.5.1	Modeling techniques . . . . .	119
3.5.2	MCP-Mod: An integrated dose-finding approach combining multiple comparisons and modeling . . . . .	121
3.5.3	Asthma dose-finding trial . . . . .	125
<b>4</b>	<b>Analysis of Multiple Endpoints in Clinical Trials</b>	<b>131</b>
	<i>Ajit C. Tamhane and Alex Dmitrienko</i>	
4.1	Introduction . . . . .	131
4.2	Inferential goals . . . . .	133
4.2.1	At-least-one procedures . . . . .	133
4.2.2	Global procedures . . . . .	134
4.2.3	All-or-none procedures . . . . .	135
4.2.4	Superiority-noninferiority procedures . . . . .	135
4.3	At-least-one procedures . . . . .	136
4.3.1	Procedures based on univariate $p$ -values . . . . .	136
4.3.2	Parametric and resampling-based procedures . . . . .	143
4.4	Global testing procedures . . . . .	147
4.4.1	Normal theory model . . . . .	147
4.4.2	OLS and GLS procedures . . . . .	148
4.4.3	Likelihood ratio and other procedures . . . . .	153
4.4.4	Procedures for individual endpoints . . . . .	155
4.5	All-or-none procedures . . . . .	157
4.6	Superiority-noninferiority procedures . . . . .	159
4.7	Software implementation . . . . .	163
<b>5</b>	<b>Gatekeeping Procedures in Clinical Trials</b>	<b>165</b>
	<i>Alex Dmitrienko and Ajit C. Tamhane</i>	
5.1	Introduction . . . . .	165
5.2	Motivating examples . . . . .	166
5.2.1	Clinical trials with serial gatekeepers . . . . .	167
5.2.2	Clinical trials with parallel gatekeepers . . . . .	168
5.2.3	Clinical trials with tree-structured gatekeepers . . . .	170



5.3	Serial gatekeeping procedures . . . . .	171
5.3.1	General serial gatekeeping framework . . . . .	172
5.3.2	Serial gatekeeping procedures with a single decision-making branch . . . . .	173
5.3.3	Serial gatekeeping procedures with multiple decision-making branches . . . . .	175
5.3.4	Other serial gatekeeping procedures . . . . .	176
5.4	Parallel gatekeeping procedures . . . . .	177
5.4.1	General parallel gatekeeping framework . . . . .	177
5.4.2	Multistage parallel gatekeeping procedures . . . . .	178
5.4.3	Cardiovascular clinical trial example . . . . .	182
5.5	Tree gatekeeping procedures . . . . .	186
5.5.1	General tree gatekeeping framework . . . . .	186
5.5.2	Combination-therapy clinical trial example . . . . .	188
5.6	Software implementation . . . . .	189
<b>6</b>	<b>Adaptive Designs and Confirmatory Hypothesis Testing</b>	<b>193</b>
	<i>Willi Maurer, Michael Branson, and Martin Posch</i>	
6.1	Introduction . . . . .	193
6.2	Basic principles and methods of error rate control . . . . .	196
6.2.1	Introduction . . . . .	196
6.2.2	Causes of multiplicity and bias in adaptive designs . . . . .	196
6.2.3	Repeated hypothesis testing at interim analyses in group sequential designs . . . . .	197
6.2.4	Adaptation of design features: Sample size . . . . .	200
6.2.5	Multiple hypothesis selection and testing in adaptive designs . . . . .	204
6.3	Principles of adaptive testing procedures . . . . .	210
6.3.1	Combination test principle . . . . .	211
6.3.2	Conditional error principle . . . . .	212
6.4	Adaptive multiple testing procedures . . . . .	214
6.4.1	Adaptive closed tests . . . . .	214
6.4.2	Estimation . . . . .	218
6.5	Case studies . . . . .	220
6.5.1	Treatment selection design . . . . .	220
6.5.2	Subpopulation selection design . . . . .	229
6.6	Discussion . . . . .	235
<b>7</b>	<b>Design and Analysis of Microarray Experiments for Pharmacogenomics</b>	<b>239</b>
	<i>Jason C. Hsu, Youlan Rao, Yoonkyung Lee, Jane Chang, Kristin Bergsteinsdottir, Magnus Karl Magnússon, Tao Wang, and Eiríkur Steingrímsson</i>	
7.1	Potential uses of biomarkers . . . . .	239
7.2	Clinical uses of genetic profiling . . . . .	240

7.3	Two stages of pharmacogenomic development . . . . .	241
7.4	Multiplicity in pharmacogenomics . . . . .	243
7.4.1	Multiplicity of genes . . . . .	244
7.4.2	Multiplicity of subgroups . . . . .	248
7.5	Designing pharmacogenomic studies . . . . .	250
7.5.1	Control for platform change using external reference sample . . . . .	250
7.5.2	Design to discover group differences . . . . .	250
7.5.3	Permutation tests may not control error rates . . . . .	254
7.6	Analyzing microarray data by modeling . . . . .	255
7.7	A proof of concept experiment . . . . .	257
7.8	Software implementation . . . . .	262
7.8.1	SAS procedures . . . . .	262
7.8.2	R Functions . . . . .	263
<b>Bibliography</b>		<b>265</b>

---

# *Preface*

Methods for addressing multiplicity issues have attracted much attention in the statistical literature over the past twenty years. Recent developments in this area include new approaches to defining false-positive error rates and new classes of multiple comparison procedures. The books by Hochberg and Tamhane (1987), Westfall and Young (1993) and Hsu (1996) introduce general statistical theory of multiple comparisons. The main goal of this book is to review and summarize the fast growing area of multiple comparison research with emphasis on pharmaceutical applications.

The book consists of seven chapters that describe important multiplicity problems encountered in pre-clinical and clinical trial settings. Each chapter provides a detailed overview of methodological issues in multiple testing with emphasis on recently developed approaches not covered in other books. Case studies based on pre-clinical experiments and clinical trials are used to help the reader quickly learn popular multiple testing methods and apply them to real-life problems. The book offers advice from subject matter experts and reviews relevant regulatory guidelines. It also provides useful information for practitioners by emphasizing technical details and implementation of the statistical methods using popular statistical software, including SAS and R.

This book is aimed mainly at biostatisticians involved in pre-clinical and clinical trial research. However, the individual chapters include high-level introductory material to make them accessible to a broad audience of pharmaceutical researchers, including drug discovery scientists, medical scientists and regulatory scientists.

## **Outline of the book**

**Chapter 1 (Multiplicity Problems in Clinical Trials: A Regulatory Perspective).** This chapter gives a broad introduction to different types of multiplicity problems that commonly arise in confirmatory controlled clinical trials. It focuses on multiplicity induced by multiple endpoints as well as other multiple comparison problems, including problems encountered in trials with multiple dose-control comparisons, trials with multiple subgroups, trials with an active control, etc. The chapter also discusses multiplicity considerations for safety endpoints and multiplicity concerns for several special situations.

**Chapter 2 (Multiple Testing Methodology).** This chapter gives an

overview of concepts and principles that play a central role in multiple testing. This includes definitions of error rates and popular testing principles (closure and partitioning principles). The chapter also introduces multiple testing procedures widely used in pre-clinical and clinical studies, including procedures based on univariate  $p$ -values, parametric procedures and resampling-based procedures. These topics provide a foundation for the pharmaceutical applications considered in the subsequent chapters.

**Chapter 3 (Multiple Testing in Dose-Response Problems).** This chapter provides an overview of statistical methods for analyzing clinical dose response studies comparing several dose levels with a control. The emphasis is on efficacy evaluation but most methods can be directly applied to safety or combined efficacy/safety evaluation. Three distinct classes of methods are discussed. This chapter begins with a review of trend tests to detect an overall dose response effect. Next, problems of finding the minimum effective dose and the maximum safe dose using multiple hypotheses testing methods are considered. Finally, the chapter discusses fitting models to dose response curves, and combining the modeling information with hypotheses testing approaches to obtain more powerful hybrid multiple comparison procedures. Illustrative numerical examples are given and available software is mentioned.

**Chapter 4 (Analysis of Multiple Endpoints in Clinical Trials).** This chapter provides an overview of statistical methods for analyzing multiple endpoints in clinical trials for comparing a treatment with a control (placebo). Four classes of methods are discussed in the chapter:

- Union-intersection procedures and other multiple testing procedures for demonstrating the treatment's superiority on at least one endpoint.
- Global procedures for demonstrating the combined effect of the treatment on all endpoints.
- Intersection-union procedures for demonstrating the treatment's superiority on all endpoints.
- Hybrid superiority-noninferiority procedures for establishing the treatment's superiority on at least one endpoint and noninferiority on all other endpoints.

Procedures based on  $p$ -values, normal theory and resampling are discussed. Illustrative examples are given from recent clinical trials.

**Chapter 5 (Gatekeeping Procedures in Clinical Trials).** This chapter describes a class of procedures, called gatekeeping procedures, for testing hierarchically ordered hypotheses. Such hypotheses commonly arise in clinical trials when dealing with multiple endpoints, dose-control comparisons and subgroup analyses. The chapter reviews three main classes of gatekeeping procedures (serial, parallel and tree-structured gatekeeping procedures) that

control the overall Type I error rate and efficiently account for the hierarchical structure of multiple objectives.

**Chapter 6 (Adaptive Designs and Confirmatory Hypothesis Testing).** This chapter provides an overview of statistical methods for the design and analysis of adaptive designs and related confirmatory hypotheses testing problems. The chapter starts with a discussion of several causes of multiplicity and bias in adaptive designs, followed by a brief review of repeated hypothesis testing at interim analyses leading to group-sequential designs as well as common blinded and unblinded sample size adjustment methods. The major focus of the chapter is on multiple hypothesis selection and testing in adaptive designs based on the closure principle applied to combination tests or conditional error rate functions. Applications of these methods include adaptive treatment or subgroup selection at an interim analysis, which are illustrated by two real case studies.

**Chapter 7 (Design and Analysis of Microarray Experiments for Pharmacogenomics).** Pharmacogenomics is the co-development of a drug that targets a subgroup of patients and a device that predicts whether a patient is in the subgroup of responders to the drug. Such a development involves a training study, followed by a validation study if warranted. This chapter discusses the design of pharmacogenomic studies based on established statistical principles and describes the analysis of data collected in these studies in a way that takes the multitude of multiplicity issues into account. Both aspects are critical to the success of pharmacogenomic development. A proof of concept experiment is used to show how proper design and analysis can smooth the path from discovery to clinical use.

## Web Site

The code and data sets included in this book are available on the book's Web site at <http://www.multipert.com>.

## Acknowledgments

We would like to thank the reviewers who have provided valuable comments on selected chapters in the book: Timothy Costigan (Eli Lilly and Company), Olivier Guilbaud (AstraZeneca), George Kordzakhia (U.S. Food and Drug Administration), Walter Offen (Eli Lilly and Company), Katherine Pollard (University of California at San Francisco).

We would also like to thank the book's acquisitions editor, David Grubbs, for his support and his work on this book publishing project.

Alex Dmitrienko, Eli Lilly and Company, U.S.A.

Ajit C. Tamhane, Northwestern University, U.S.A.

Frank Bretz, Novartis, Switzerland; Hannover Medical School, Germany.

---

## *List of Contributors*

This book is based on a collaborative effort of 20 statisticians, scientists and clinicians from pharmaceutical industry, academia and regulatory agencies:

**Kristin Bergsteinsdottir**, Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Iceland, Reykjavík, Iceland.

**Michael Branson**, Director, Clinical Information Sciences, Novartis Pharma AG, Basel, Switzerland.

**Frank Bretz**, Biometrical Fellow, Clinical Information Sciences, Novartis Pharma AG, Basel, Switzerland; Adjunct Professor, Hannover Medical School, Hannover, Germany.

**Jane Chang**, Associate Professor, Department of Applied Statistics and Operations Research, Bowling Green State University, Bowling Green, Ohio, U.S.A.

**Alex Dmitrienko**, Research Advisor, Global Statistical Sciences, Eli Lilly and Company, Indianapolis, Indiana, U.S.A.

**Jason C. Hsu**, Professor, Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A.

**Mohammad Huque**, Division Director, Division of Biometrics IV/OB/OTS/CDER, Food and Drug Administration, Silver Spring, Maryland, U.S.A; Adjunct Professor, J. P. Hsu College of Public Health, Georgia Southern University, Statesboro, Georgia, U.S.A.

**Yoonkyung Lee**, Associate Professor, Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A.

**Magnus Karl Magnússon**, Landspítali-University Hospital, Reykjavík, Iceland.

**Willi Maurer**, Senior Statistical Consultant, Clinical Information Sciences, Novartis Pharma AG, Basel, Switzerland.

**José Pinheiro**, Senior Biometrical Fellow, Clinical Information Sciences, Novartis Pharmaceuticals, East Hanover, New Jersey, U.S.A.

**Martin Posch**, Associate Professor, Section of Medical Statistics, Medical University of Vienna, Vienna, Austria.

**Youlan Rao**, Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A.

**Joachim Röhmel**, Professor, Institute for Prevention Research and Social Medicine, Bremen, Germany.

**Eiríkur Steingrímsson**, Professor, Department of Biochemistry and Molecular Biology, Faculty of Medicine, University of Iceland, Reykjavík, Iceland.

**Ajit C. Tamhane**, Senior Associate Dean and Professor of IEMS, McCormick School of Engineering and Applied Science, Northwestern University, Evanston, Illinois, U.S.A.

**James Troendle**, Senior Investigator, Division of Epidemiology, Statistics and Prevention Research, National Institutes of Health, Bethesda, Maryland, U.S.A.

**Tao Wang**, Pfizer Inc, New London, Connecticut, U.S.A.

**Peter H. Westfall**, Professor, Information Systems and Quantitative Sciences, Texas Tech University, Lubbock, Texas, U.S.A.

**Brian L. Wiens**, Director, Development, Alcon Laboratories, Inc., Fort Worth, Texas, USA.

# Chapter 1

---

## *Multiplicity Problems in Clinical Trials: A Regulatory Perspective*

**Mohammad Huque**

*U.S. Food and Drug Administration*

**Joachim Röhmel**

*Bremen Institute for Prevention Research and Social Medicine*

---

### **1.1 Introduction**

Confirmatory controlled clinical trials, also known as Phase III clinical trials, when successful, are significant achievements in medical research as they provide evidence that new treatments (e.g., test drugs or other types of interventions) studied in these trials are clinically effective in treating targeted diseases, and are also safe as far as judgment is possible on taking into account the limited number of exposed patients. Unfortunately, many such trials fail and are unable to show that new treatments studied in these trials are better than placebo. This is surprising because Phase III clinical trials are designed and conducted after the so-called Phase II trials, which are supposed to have already shown promising evidence of treatment efficacy and safety. There can be several reasons for such failures. For example, certain weaknesses in the primary endpoints of a trial can jeopardize the success of a trial, e.g., if these endpoints are not objective, or are not validated, or are not in line with the mechanisms of actions of the treatment. A trial can also fail because of poor planning or disregarding multiplicity issues with respect to multiple endpoints and multiple comparisons.

Clinical trials generally pose multiple questions in the form of hypotheses whose evaluations involve multiple comparisons and tests for multiple endpoints. In testing a single hypothesis, a statistical test in the absence of a treatment effect can lead to a positive conclusion in favor the treatment effect just by chance. Such an error in the testing of hypotheses terminology is

---

Views expressed in this chapter are the personal views of the authors and not necessarily of any regulatory agency such as the U.S. Food and Drug Administration.



known as a false positive error or a Type I error. When multiple hypotheses are tested without an appropriate adjustment, this error can become excessive. In other words, the familywise error rate (FWER) defined in Section 2.2 can become inflated. This situation can then lead to an easy approval of an ineffective treatment. Therefore, it is important that trials control this error probability at a prespecified level through appropriate design and analyses strategies that are prospectively planned.

This introductory chapter discusses multiplicity problems that commonly arise in confirmatory controlled clinical trials, and also briefly reviews new more efficient methods for dealing with these problems. These methods are described in more detail in subsequent chapters. The types of multiplicity problems that arise in clinical trials depend on how these trials are designed for assessing clinically meaningful benefits of treatments under investigation, given that these trials may collect data on multiple endpoints at different time points on patients randomized to test and control groups at different dose levels. Some trials may also perform multiple interim analyses during the course of the trial and change some design features adaptively based on the results of interim analyses. This creates additional layers of multiplicity problems. Some trials may also be multiregional for which treatment benefit evaluations may require addressing regional differences, e.g., differences in clinical practice, patient ethnicity and other clinical and biological factors. These trials may pose challenging design and multiplicity problems, when in the absence of consistency of treatment effects, evidence of treatment benefits may be sought for specific regions or sub-populations.

With the above background, Section 1.2 introduces the most common multiplicity problems in confirmatory trials. This includes multiplicity problems arising in trials with multiple endpoints when treatment efficacy conclusions are made through the so-called clinical “win” criteria. A win criterion is basically a clinical criterion that defines what sort of endpoint results need to be observed in a trial for a specific disease indication in order to conclude that the treatment provides clinically meaningful treatment benefits. Further, this section reviews multiple comparison problems in trials with a single primary endpoint, e.g., trials with multiple doses compared to a common control or trials with multiple subgroups, and more advanced multiplicity problems, e.g., trials with ordered multiple primary and secondary endpoints. Note that this chapter does not deal with special multiplicity topics, such as multiple comparison problems in trials utilizing adaptive designs. These problems are discussed in Chapter 6.

Here is an outline of the remaining sections. Section 1.3 discusses methods that can be used to reduce the burden of multiplicity in clinical trials, including the use of composite endpoints. Section 1.4 addresses multiplicity concerns for a few special situations that frequently occur in clinical trials such as the use of multiple patient populations. Section 1.5 reviews multiplicity issues for safety endpoints. Finally, Section 1.6 provides some concluding remarks.

Each main section, where applicable, provides necessary motivating ex-

amples and classification of the problems at hand before providing a more detailed overview of relevant multiple comparison problems.

---

## 1.2 Common multiplicity problems in clinical trials

This section begins with a review of issues related to the analysis of multiple endpoints in confirmatory clinical trials. The topic of multiple endpoints is discussed in greater detail in Chapter 4. Clinical trials generally classify the endpoints into primary, secondary and exploratory types. Primary endpoints address primary objectives of the trial. They are usually few but are clinically most relevant to the disease and the treatment under study. They assess the main clinical benefits of the treatment. This is usually done through one or more clinical “win” criteria. Examples of such criteria and discussion of multiplicity problems they pose are given in Section 1.2.1. Section 1.2.2 introduces a consistency-based approach to the analysis of multiple endpoints. Section 1.2.3 discusses a specific type of win criterion stating that all primary endpoints individually must show statistically significant treatment benefit. Primary endpoints of this type are usually referred to as co-primary endpoints.

Secondary endpoints characterize extra benefits of the treatment under study after it has been demonstrated that the primary endpoints show clinically meaningful benefits of the treatment. D’Agostino (2000) and Moyé (2003) give detailed classifications of secondary endpoints. O’Neill (1997) supported the idea that “secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance.” Section 1.2.4 discusses multiplicity adjustment issues for secondary endpoints.

Exploratory endpoints, for reasons of their own, are usually not prospectively planned and are generally not rigorously evaluated like primary and secondary endpoints. These endpoints are used in treatment comparisons and also unplanned subgroup analyses with an exploratory (e.g., hypothesis generating) purpose. In certain situations, their results can be useful in designing future new trials. However, they are not useful for confirmatory purpose, as there are no valid ways for controlling the FWER for their results. Results of such analyses have the potential of being observed by chance alone.

Further, Sections 1.2.5 to 1.2.8 review other common multiplicity problems in confirmatory clinical trials. Section 1.2.5 discusses multiplicity problems in dose-response and dose-control comparison trials. Section 1.2.6 describes approaches to multiple comparison problems in trials with planned subgroup analyses, e.g., subgroups defined using certain validated biomarker classifiers that are expected to enjoy better treatment efficacy and/or safety. Sections 1.2.7 and 1.2.8 cover multiplicity issues in drug combination and gold standard trials, respectively.

### 1.2.1 Clinical win criteria with primary endpoints

At times, win criteria are also called “clinical decision rules” for determining clinically meaningful treatment efficacy. They simply define how a positive clinical decision regarding the effectiveness of a test treatment in a trial is going to be reached. The criteria are defined relative to one or more relevant clinical primary endpoints in the setting of comparing one or more doses of test and control treatments. They represent a clinically acceptable way of assessing the effectiveness of a treatment for a disease and patient population under consideration. If statistically significant results are achieved satisfying such criteria, then they can justify meaningful clinical benefits of the study treatment. Although this section focuses on multiple endpoints, the same principles and arguments can be applied to clinical trials with multiple objectives, for example, trials with more than two treatments or trials with inference at more than one time point. Some examples of possible win scenarios with multiple endpoints are as follows:

Example 1. One specified primary endpoint needs to be significant.

Example 2. Given  $m \geq 2$  specified primary endpoints, at least one needs to be statistically significant.

Example 3. Given  $m \geq 2$  specified primary endpoints, all need to be statistically significant.

Example 4. Given three specified primary endpoints E1, E2 and E3, either E1 needs to be statistically significant or both E2 and E3 need to be statistically significant.

Example 5. Given three specified primary endpoints E1, E2 and E3, either (both E1 and E2) or (both E1 and E3) need to be statistically significant.

Example 6. One of the two specified endpoints must be statistically significant and the other one needs to be at least marginally significant.

Example 7. One of the two specified endpoints must be statistically significant and the other one needs to show non-inferiority.

Example 8. Several endpoints are in a hierarchy that specifies the order in which they are to be evaluated, and significance of all preceding endpoints is required in order to proceed to the next one.

Example 9. Of the  $m$  specified primary endpoints, at least  $k$  endpoints need to be statistically significant and the remaining  $m - k$  endpoints trend in the right direction.

Example 10. There are complex logical relationships among specified primary and secondary endpoints, e.g., secondary null hypotheses need to be tested only if all primary null hypotheses are rejected or at least one primary null hypothesis is rejected.

Examples 1, 3 and 8 do not require any adjustment for multiplicity. For these three examples, each test can be performed at the same significance level of  $\alpha$ . Example 1 is the single primary endpoint case, as such, there is no multiplicity issue for this case. Example 3 is the case of co-primary endpoints where all endpoints must show treatment benefit. Testing for treatment benefits for co-primary endpoints follows the intersection-union principle which is discussed later in this section and also in Section 2.3.2. In Example 8, the region of rejection of each test falls within the region of rejection of the previous test whose size is  $\alpha$ . Therefore, in this sequential testing scheme (see Section 2.6.3) the FWER always remains bounded above by  $\alpha$ . Example 8 type of decision strategy can be useful in getting an earlier onset claim of a treatment benefit in addition to treatment benefits for the usual clinically required onset and durability endpoints. For example, a hair growth trial may test first in sequence the 12th and 15th month endpoints for the usual onset and durability claims of the treatment effects. After this claim is achieved, then one can go back and test the earlier 9th month endpoint for an earlier benefit of the treatment for an additional claim.

Example 2 is the case where each primary endpoint can on its own characterize a clinically meaningful benefit of the treatment under study. Therefore, a win in at least one of the primary endpoints is a win for the trial and the treatment benefits can be labeled as such depending on which endpoints show statistically significant treatment benefits after multiplicity adjustments. Testing for this type of win criterion follows the union-intersection principle which requires multiplicity adjustments for FWER control. This testing principle is covered in Section 2.3.1. For example, a cardiovascular trial may achieve its primary objective for a clinical benefit, if it establishes the efficacy for the study treatment either for all-cause mortality or for MI (myocardial infarction) or for stroke. Thus, the win criterion for this example is to win for efficacy on at least one of the three endpoints. However, in a trial like this when treatment effect sizes are expected to be small but clinically meaningful, an alternative approach is to define a composite endpoint as the primary endpoint. This and the caveats in creating composite endpoints are addressed in Section 1.3.

In Example 4, the null hypothesis is an intersection of two null hypotheses. The first one is a simple null hypothesis that there is no treatment effect for the endpoint E1. The second one is a union null hypothesis that there is no treatment effect for at least one of the two endpoints E2 and E3. The intersection of these two null hypotheses can be tested using the Bonferroni procedure introduced in Section 2.6.1. For example, test E1 at the significance level of  $\alpha_1$  and test E2 and E3 each at the significance level of  $\alpha - \alpha_1$ . An alternative approach would be to test E1 first at a reduced significance level of  $\alpha_1$  (e.g.,  $\alpha_1 = 0.025$ ) and if the result is statistically significant at this level, then test each E2 and E3 at the full significance level of  $\alpha$  (e.g.,  $\alpha = 0.05$ ), else, test each E2 and E3 at the reduced significance level of  $\alpha - \alpha_1$ . This latter test is the fallback procedure (see Section 2.6.4). Both approaches will control

FWER in the strong sense, the latter one with a better power for situations when there is a treatment effect for E1.

In Example 5, (E1, E2) and (E1, E3) are pairwise co-primary. This win criterion is equivalent to win for E1 and win for either E2 or E3. In this case, E1 is considered the most relevant endpoint for treatment benefit, but a statistically significant result for E1 alone is considered insufficient to characterize the clinical benefit of the treatment. The clinical decision rule requires an additional statistically significant result in favor of the treatment in at least one of the other two endpoints E2 and E3. In this case, the null hypothesis for the most relevant endpoint E1 and the intersection null hypothesis for the other two softer endpoints E2 and E3 are hierarchically ordered – there is no need for testing for endpoints E2 and E3 unless the null hypothesis for the endpoint E1 is rejected first. Therefore, one can test E1 at the full significance level of  $\alpha$ , and if this result is statistically significant then proceed to test for E2 and E3 using a FWER controlling method (e.g., Hochberg or Holm procedures described in Section 2.6) at level  $\alpha$ .

Example 6 is the case of two co-primary endpoints, where one endpoint must show convincing evidence of efficacy, and the other one at least marginally. Such a relaxed evidence criterion for a pair of co-primary endpoints is meaningful only if it is clinically acceptable in providing adequate evidence of efficacy for the treatment and the disease under study. Methods for relaxing the evidence criteria to improve the power of co-primary endpoint tests are discussed later in this section and in Section 4.5.

Instead of relaxing the evidence criterion for one endpoint one could set (non-inferiority) margins and require superiority in the more important endpoint and at least non-inferiority for the less important endpoint (Example 7). Because there are two possibilities to win, a multiplicity adjustment needs to be applied to control the FWER (Tamhane and Logan, 2004; Röhmle et al., 2006; Logan and Tamhane, 2008). Superiority/non-inferiority procedures for multiplicity problems of this type are discussed in Section 4.6.

The decision rule in Example 9 has been used in a few arthritis trials and requires appropriate multiplicity adjustments. An arthritis trial generally includes four key endpoints, namely, joint counts for tenderness and swelling, and physician and patient global assessments. As all these endpoints are largely impacted by the condition of the patient's joints, clinical expectations have been that all these four endpoints should show some evidence of treatment benefits. However, because of small treatment effect sizes in some of these endpoints that can result in lack of power of the tests, a clinical decision rule used was to win in at least three out of the four endpoints with one endpoint at least trending in the right direction. However, recent arthritis trials, instead of testing for multiple efficacy endpoints test for a single responder endpoint. This endpoint known as ACR20 combines patient outcomes from seven efficacy endpoints to a single responder endpoint based on a clinical decision rule that determines whether a patient has responded to the treatment

or not. Felson et al. (1995) gave the definition of this endpoint with some validation results.

Example 10 encompasses multiplicity problems of clinical trials that are more complex. They require setting up hybrid multi-stage testing procedures that combine serial and parallel gatekeeping testing strategies and take into account applicable logical restrictions. Problems of this type are discussed later in this section and in Chapter 5.

It should be apparent from the above examples that a null hypothesis for a given clinical decision rule can be complex. A clinical decision rule for winning basically defines an alternative hypothesis space and its complement is then the null hypothesis space. Its statistical testing, when done efficiently with adjustment for multiplicity, can provide clear evidence of clinically relevant benefits of the study treatment.

### **1.2.2 Consistency-ensured analysis of multiple endpoints**

Clinical trials may include more than one clinically important endpoint, each with ability to characterize clinically meaningful benefits of the treatment under study. However, to keep the study size and/or its duration feasible, such a trial may designate only one of these clinically important endpoints as primary leaving others with unknown power properties. Trials of this type can lead to unsettling situations, if the result for the designated primary endpoint is not statistically significant but the result for another alternative clinically important endpoint is strongly positive. Such unsettling situations can arise because of unresolved issues regarding:

- The alternative endpoint, although it produces a strongly positive result, is not a prospectively planned alternative primary with proper  $\alpha$  adjustments for evaluating its efficacy.
- The designated primary endpoint result is much weaker than expected or is in the opposite direction causing difficulties in interpreting study findings.

Clinical trial literature reports several trials of this type that had difficulties in the interpretation of study results. An example is the carvedilol trials which were discussed by the FDA Cardiovascular and Renal Drugs Advisory Committee in May 1996. Details about this can be found in Packer et al. (1996) and Fisher and Moyé (1999). In this case, the experimental treatment failed to show a statistically significant result for the planned primary endpoint, namely change in the exercise tolerance, but the results revealed a striking reduction in mortality. Moyé (2003), Moyé and Baraniuk (2007) and Li and Mehrotra (2008) mentioned several other cases of the type where trials were powered for a designated primary endpoint but had interests in testing for another alternative clinically important endpoint as well. The issue of specifying an alternative endpoint can be resolved through a pre-planned

statistical testing strategy that controls the FWER adequately for evaluating both the designated and the alternative endpoint results. Literature has recently introduced several approaches that address this problem, including the prospective alpha allocation scheme (PAAS), nonparametric and parametric fallback methods defined in Chapter 2. The PAAS method has some power advantage over the weighted Bonferroni approach. The nonparametric fallback method for ordered hypotheses gains additional power by allocating a higher  $\alpha$  level to a hypothesis if the hypothesis earlier in the sequence is rejected. Similarly, the parametric fallback method attempts to gain additional power over the regular fallback method by taking into account the correlation among the endpoints. However, although the fallback method and its extensions are attractive in saving  $\alpha$  for testing additional hypotheses, they do not go beyond the PAAS approach for increasing the chance of a positive trial if a trial does not establish efficacy on the first endpoint. Also, none of these approaches address the issue of interpretation of the study findings when the results for the designated primary and the alternative endpoint are either far apart from each other or are in opposite directions. Recent statistical literature has introduced some methods that weigh the evidence of the two clinically important endpoints of a trial for concluding treatment efficacy. The adaptive alpha allocation approach, called the 4A method by Li and Mehrotra (2008), is an interesting start in that it incorporates the result of the primary endpoint in making inference about the alternative clinically important endpoint. However, further research is needed for better understanding the properties of this method for trials when clinically important endpoints are likely to be correlated. Note that this method mimics the PAAS but the allocated  $\alpha$  for testing the second hypothesis is calculated adaptively in a pre-specified manner based on the observed  $p$ -value for the first endpoint, so that a weaker result in the primary endpoint requires a much stronger result for the alternative endpoint.

On the other hand, Song and Chi (2007) and Alosch and Huque (2007, 2009) proposed approaches for subgroup analyses that require establishing a certain degree of evidence of efficacy on the primary analysis (i.e., for the total population) before proceeding to test for its subgroup. Huque and Alosch (2009) applied this concept of “consistency” in testing for the designated primary and its alternative endpoint.

### 1.2.3 Co-primary endpoints

Many trials characterize clinically meaningful efficacy of a new treatment through a single primary efficacy endpoint. However, trials for certain diseases do this through multiple primary efficacy endpoints requiring a statistically significant benefit of the new treatment on each of these endpoints. In clinical trials terminology such primary endpoints are usually referred to as co-primary endpoints. A non-significant result in any of the specified co-primary endpoints would then lead to a non-win scenario for the trial. For example, Alzheimer’s trials in mild-to-moderate disease generally include ADAS-Cog and CIBIC

(Clinician's Interview Based Impression of Change) endpoints as co-primary. The ADAS-Cog endpoint measures patients' cognitive functions while the CIBIC endpoint measures patients' deficits of activities of daily living. For a claim of clinically meaningful treatment benefit for this disease, clinicians generally require demonstration of statistically significant treatment benefit on each of these two primary endpoints.

Clinical considerations, such as what benefits a patient in a clinically meaningful way, generally drive the inclusion of two or more relevant primary endpoints in a trial as co-primary. These endpoints are generally used for diseases that manifest in multiple symptoms. All relevant symptoms of the disease must be controlled for a treatment to be viable for such a disease. For example, patients with migraine experience severe headache, typically associated with photophobia, phonophobia, and nausea/vomiting. If a migraine trial shows evidence of treatment benefit only for the endpoint of "headache" and not for other endpoints, then it may earn claim of treatment benefit for headache, but it may fail to do so for the treatment of "migraine." Clinically meaningful treatment benefit claim for migraine, at least in the U.S. regulatory setting, usually requires that the treatment, besides being safe, is effective in relieving all the necessary clinical symptoms of migraine including headache. Additional clinical trial examples of co-primary endpoints can be found, for example, in several CHMP guidelines available at

<http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm>

and in Offen et al. (2007).

The use of co-primary endpoints in a trial puts extra burden on the trial as it raises the bar for the evidence of efficacy - one must show evidence of efficacy in more than one specified primary endpoints. This generally causes enlarging the size of the trial on carefully assessing that each co-primary endpoint has adequate sensitivity for detecting a desired treatment benefit. Section 4.5 evaluates increases in the sample size in trials with co-primary endpoints in relation to that for single primary endpoint trials. However, despite this sample size concern, many co-primary endpoint trials are successfully done. A reason for this is that the endpoint treatment effect sizes for these trials are fairly reasonable, often 0.3 standard deviations or greater (see Huque et al., 2008). In addition, some of these trials being symptomatic trials are not difficult to conduct with respect to patient enrollment and endpoint ascertainment. [Table 1.1](#) gives some idea of the sample sizes for trials with 2 to 4 co-primary endpoints in relation to single endpoint trials, when effect sizes for co-primary endpoints are in the range 0.2 to 0.4 standard deviation. Note that this table is different than that given in Section 4.5 in that it gives some idea about the size of the trial in terms of the endpoint effect sizes. This table shows that the co-primary endpoint trials can be feasible with respect to sample sizes if the endpoint treatment effect sizes are about 0.25 standard deviation or better. Note that trials with other types of win criteria that require multiplicity ad-



**TABLE 1.1:** Sample sizes per treatment arm in a trial with  $m$  co-primary endpoints ( $\alpha = 0.025$ , one-sided tests, power = 80%). The test statistics follow a multivariate normal distribution with equal pairwise correlations and common effect size.

Effect size	Correlation	Single endpoint trial	Trial with $m$ co-primary endpoints		
			$m = 2$	$m = 3$	$m = 4$
0.20	0.0	393	516	586	636
	0.4		497	556	597
	0.8		458	494	518
0.25	0.0	252	330	375	407
	0.4		318	356	382
	0.8		293	316	332
0.30	0.0	175	230	261	283
	0.4		221	247	266
	0.8		204	220	230
0.40	0.0	99	129	147	159
	0.4		124	139	150
	0.8		115	124	130

justments also decrease power and require sample size increase in comparison to trials with a single endpoint.

Statistically, the null hypothesis for testing co-primary endpoints is a union null hypothesis and the alternative an intersection hypothesis. A test of such a null hypothesis is sometimes referred to in the statistical literature as an intersection-union test (see Section 2.3.2). This test does not inflate the Type I error rate, and as such, there is no downward adjustment of  $\alpha$  for this case, i.e., each co-primary endpoint test can be performed at the same significance level of  $\alpha$ . If each of the  $m$  co-primary endpoints is tested at the significance level of  $\alpha$  then the maximum Type I error rate for testing  $m$  co-primary endpoints is the same as  $\alpha$ . However, testing for co-primary endpoints inflates the probability of the Type II error, and thus, reduces the power of the test.

In testing for co-primary endpoints, the inflation of the Type II error rate, and consequently the reduction in the power of the test, besides depending on  $m$ , also depends on the extent of dependency between the test statistics of the co-primary endpoints. If these test statistics were independent, the power would simply be the product of the powers for single endpoint tests, with Type II error probability as one minus this power. If the test statistics were perfectly correlated, and the treatment effects sizes (per unit standard deviation) for the co-primary endpoints were homogeneous, then there would be no inflation of the Type II error probability. As the correlation between test-statistics can fall in between the two extremes, so can be the power of the co-primary endpoint tests. Therefore, in designing clinical trials with co-primary endpoints, one must also consider the dependency between the test statistics for addressing the Type II error probability.

In testing for co-primary endpoints, treatment effect sizes also impact the power of the test. In a trial, one may not expect the same treatment effect size for all co-primary endpoints. Some endpoints, such as low event endpoints, can have relatively small effect sizes in comparison to other endpoints, causing the trial to be powered with respect to the low yield endpoints. However, in some such situations, this difficulty can be overcome to some extent by adequately enriching the trial, or alternatively, low yield endpoints, if they are likely to exhibit similar treatment effects, by combining them to a composite endpoint, thus reducing the dimensionality in the co-primary endpoint testing. Such a composite endpoint can be acceptable if it is clinically relevant for the disease under study and has regulatory and scientific acceptance. Section 1.6 addresses considerations for composite endpoints.

An important consideration in designing a clinical trial with co-primary endpoints is to be sure that the compound under study is such that it has the ability through its mechanism of actions to target all the co-primary endpoints of the trial. Before launching such a trial, scientists usually validate this assertion through animal studies, early phase human trials (such as proof-of-concept trials) and on synthesizing historical trials of similar compounds. A trial with co-primary endpoints will surely fail if an endpoint in the co-primary set has no efficacy sensitivity for the given compound.

#### **1.2.4 Secondary endpoints**

As indicated at the beginning of this section, the role of secondary endpoints in confirmatory clinical trials is considered different than that of the primary endpoints. Primary endpoints address primary objectives of the trial. On the other hand, secondary endpoints have a number of important functions at levels different than those of primary endpoints. Details of these functions of secondary endpoints can be found in D'Agostino (2000) and in Moyé (2003). It is well understood that secondary endpoint hypotheses can be tested for extended treatment benefits after the primary objectives of the trial have been successfully met. However, at times, for planning purposes, a key endpoint such as mortality is called a secondary endpoint on expecting it to be a low yield endpoint. In this case, such a secondary endpoint is like a primary endpoint and is sometimes called a key secondary endpoint. Its result, after proper multiplicity adjustments, if statistically significant in favor of the treatment, can provide a persuasive evidence of a clinical benefit of the study treatment.

Different clinical trials depending on the objectives often take different approaches towards multiplicity adjustments for the analysis of secondary endpoints. These approaches vary from no adjustments to adjustments with strong control of the FWER. If the purpose of secondary endpoint analyses is to make additional claims of treatment benefits, in addition to those already established by the analyses of the primary endpoints, then multiplicity adjustments generally require a strong control of the FWER. In this regard, a useful analysis approach is the gatekeeping approach. On the other hand,

if the purpose of secondary endpoint analyses is simply to facilitate interpretations of the results of the primary endpoints, and there is no intent for additional claims of treatment benefits, then a strong control of the FWER for secondary endpoint analyses may not be necessary. However, regardless of the approach taken for the analysis of secondary endpoints in a specific trial, there is a general agreement that the secondary endpoints along with their methods of analysis should be prospectively planned, and the outcomes of these endpoints should be carefully ascertained similar to those for primary endpoints. Following are some of the approaches that seem to have emerged in the context of the analysis of secondary endpoints.

Approach 1. Control the FWER in the strong sense at a pre-specified significance level  $\alpha$  in testing both the primary and secondary endpoint families of hypotheses. Test these two families of hypotheses hierarchically, and in this regard, consider the family of primary endpoint hypotheses ahead of the family of secondary endpoint hypotheses. This can allow endpoint specific claims of treatment benefits for secondary endpoints after the primary objectives of the trial have been met. In this case, the multiplicity adjustments for the secondary endpoints can depend on the results of the primary endpoints, but normally the multiplicity adjustments for the primary endpoints should not depend on the results of the secondary endpoint tests.

Approach 2. Once the primary objectives of the trial are met, then test for the family of secondary endpoints independently of the tests for the primary endpoints, but adjust for multiplicity by controlling the FWER for the family of secondary endpoints in the strong sense. In this case, the multiplicity adjustments for the secondary endpoints do not depend on the results of the primary endpoints – all that is needed is that the primary objectives of the trial have been met. This approach can result in an inflation of the FWER under Approach 1.

Approach 3. Analyze secondary endpoints only for supporting evidence without any intention for a claim of treatment benefit for the secondary endpoints. In this case, the  $p$ -values, confidence intervals, and any other results for secondary endpoints are purely for descriptive purposes and not for drawing inferences.

### **1.2.5 Dose-response and dose-control comparisons**

Dose-response studies in clinical settings are essential for finding a dose or dose range of a treatment which is efficacious as well as safe. However, the methods and objectives of the dose-finding analysis differ for Phase II versus Phase III trials. In Phase II trials dose-finding is generally model-based. In Phase III trials, on the other hand, dose-finding is generally based on hy-

pothesis testing, requiring pre-specification of a suitable multiple comparison procedure given the trial design and the nature of claims planned for the trial.

In Phase II trials, one fits a set of suitable parametric dose response models, such as  $E_{\max}$  or logistic models and their variants, given dose-response data on a key response variable at  $m$  doses of the test treatment including placebo as a zero dose. Fitting of these models requires estimation of the parameters of the models. A working model, which is the best candidate for describing the dose response relationship, is then selected from these fitted models by performing appropriate statistical tests. This working model is then used for proof of activity of the test treatment in the dose range of the trial and for deciding which few doses should be included in the confirmatory trial. In this approach, the need for multiplicity adjustments arises for addressing uncertainty in the model selection. Bretz et al. (2005) described this approach in detail. This approach has the advantage that it incorporates clinical pharmacology model-building concepts in setting candidate models and inferences for dose selection are not confined to the target doses among the dose levels under investigation. See Section 3.5 for more information on methods that combine multiple comparison procedures with modeling techniques.

An alternative conventional approach is based on hypothesis testing. This approach treats dose as a qualitative factor with very few assumptions, if any, about the underlying dose response model. This approach is often used for identifying the minimum effective dose (MED) that is statistically significantly superior to placebo and is clinically relevant (see Tamhane et al., 1996; ICH E4, 1994). Target dose estimation is discussed in Section 3.3. This approach, though relatively robust, does not take into account the clinical pharmacology concepts that drive the selection of suitable dose response functions to be considered for a specific treatment and disease situation. The use of model-based approaches is emphasized for Phase II trials for deciding about a therapeutic dose or dose range for confirmatory trial use.

In a Phase III setting, a general design for multiple doses is that which includes two or more doses of the test treatment, placebo, and one or more doses of active control. Bauer et al. (1998) investigated this design in detail and have proposed various inferential strategies for dealing with questions of interest and multiplicity issues. Other designs are mainly special cases of this general design as follows:

- Case 1. Trials that include 2 or more doses of a test treatment and a placebo, without any active control arm, with (a) no order restriction among doses and with (b) order restriction among doses. When there is no order restriction among doses, the global test, also called the overall heterogeneity test, can be performed, for example, by one-way analysis of variance with a common unknown variance under the normality set up, contrast-based trend tests (see Section 3.2) or by the Kruskal-Wallis test under the non-parametric set up. This global test can be replaced by the standard Dunnett procedure or Dunnett-Tamhane (1991) step-down procedure for comparing different doses to placebo (see Section

2.7). However, when order restriction among doses can be safely assumed (e.g., dose effects are in non-decreasing order by dose), one can use the Bartholomew or Williams tests discussed in Section 3.2. Alternatively, one can perform a fixed-sequence procedure presented in Section 2.6.3.

Case 2. Trials with two or more doses of the test treatment and one or more doses of an active control. These trials are unable to include placebo because of ethical reasons. Such a trial design is used for many serious diseases like meningitis infection or serious cardiovascular diseases when placebo treatment can cause irreversible harm to patients. For these trials, often it is either well-known to the disease area experts or is evident from the relevant historical data that the active control treatment efficacy benefit is substantially larger than that for placebo. For this case, the efficacy goals can then be accomplished without the placebo in the trial by the statistical non-inferiority testing methods.

Case 3. Trials with three or more doses of the treatment without placebo and without active control. This type of trial is used for establishing efficacy of a new treatment when it is unethical to include placebo in the trial and also an appropriate active control is unavailable for setting the trial as a non-inferiority trial. For such a trial, an approach for establishing efficacy of a treatment is to establish its dose response, e.g., one of the higher doses of the treatment is superior to its lower dose in efficacy. The underlying assumption is that the highest dose may not be sufficiently tolerable to have adequate compliance for demonstrating efficacy. However, at least one of the higher doses is safe, tolerable and efficacious, but some lower doses may not have sufficient efficacy. Sample size requirement can be large for such a trial. One can test for the global hypothesis for an overall assessment of efficacy followed by pairwise comparisons of dose groups with proper multiplicity adjustments.

### **1.2.6 Subgroup analyses**

Subgroup analyses are quite prevalent and are considered necessary in clinical research. Commonly, the purpose is to either justify consistency of results across clinically relevant subgroups or discover large treatment differences among certain subgroups. Regulatory guidance recommends subgroup analyses by race, gender and age for pivotal comparative trials. Subgroup analyses, however, pose analytical challenges and are generally fraught with difficult inferential issues including multiplicity and lack of power issues. Lagakos (2006) wrote, “Subgroup analyses are an important part of the analysis of comparative trials. However, they are commonly overinterpreted and can lead to further research that is misguided, or worse, to suboptimal patient care.” Yusuf et al. (1991), Wang et al. (2007) and others raised similar concerns about subgroups analyses and included some helpful recommendations.

In this section, we discuss multiple comparison aspects of subgroups analysis, and emphasize that prospective planning is key to any subgroup analysis.

Often subgroup analyses are data-driven and the biological plausibility of a positive finding is argued after the data has been analyzed and results seen. Such subgroup analyses produce results that are usually false positives and not replicable. The ISIS-2 trial (ISIS-2 collaborative group, 1988) result discussions include an interesting example in this regard. Unplanned subgroup analyses may not be helpful even for hypotheses generating purposes; it can lead to misguided future research. In this situation, it is extremely difficult to know, or to be convinced of, even if somebody keeps a record of it, as to how many analyses were done prior to finding a significant positive result. Even ignoring the issue of bias for questionable comparisons, it is extremely arduous to control the Type I error rate for such analyses without knowing the correct number of analyses performed and the pre-specified adjustment method. If the investigator was honest, and, say, he did 20 analyses in searching for a significant  $p$ -value of 0.05 or less, then the Type I error for this result could be as high as 64 percent. Another key point often ignored is that in unplanned subgroup analyses, treatment groups may not be comparable because of randomization issues causing confounding and bias in the results. Consequently,  $p$ -values for treatment comparisons cannot be validly interpreted.

Frequently subgroups formed are “improper subgroups.” An improper subgroup is defined as a group of patients characterized by a variable measured after randomization and potentially affected by treatment and study procedures. For example, an improper subgroup analysis in clinical trials is the so called “completers analysis” that excludes treatment failures. In contrast, a “proper subgroup” is a group of patients characterized by a common set of “baseline” characteristics that cannot be affected by treatment, e.g., patient demographic characteristics or prognostic disease characteristics defined before randomization.

Moyé (2003) and others argued that if multiple subgroups are analyzed, then the best estimates of the treatment effects in subgroups are not the treatment effects observed in those subgroups; rather they should be shrunk to the average treatment effect for the total patient population of the trial; see Efron and Morris (1977). Low or high treatment effect sizes in subgroups can appear even under the null hypothesis of no treatment-by-subgroup interaction because of the regression to the mean phenomenon. On the other hand, a targeted subgroup within a trial may have an atypical treatment effect. This information may be available from prior clinical trials or from Phase II trials, or from biological reasoning, e.g., from histology types, as to why a targeted subgroup may have potential to show better treatment efficacy than the rest of the patients of the trial. For example, a positive outcome of a pharmacogenomic biomarker for a patient may identify that patient at randomization to be a potential responder to the given treatment.

A subgroup analysis, whether planned or unplanned, can experience a significant loss of power in detecting a treatment effect. However, this loss may

not occur for a targeted subgroup whose effect size happens to be sufficiently greater than that for the total patient population of the trial. By using a power formula due to Koch (1997), Alosch and Huque (2008) showed that for some situations when the treatment effect size for a subgroup exceeds that of the total population, the power for the treatment effect test (after proper adjustment for multiplicity) for that subgroup can exceed that for the total population treatment effect. This observation, and the previous work by Simon and Maitournam (2004) and Maitournam and Simon (2005), opens the door for designing trials for testing for treatment effects for a planned targeted subgroup for which one expects a much larger size treatment effect than that for the total patient population. Traditional randomized trials are usually not well-planned for subgroup analyses. Such an approach may be adequate if the study treatment effect is expected to be homogeneous across subgroups. However, if this treatment effect is expected to be heterogeneous, such trials may miss identifying subgroups of patients which are most likely to benefit by the study treatment.

Composite nature of Phase III trials with respect to patient characteristics is well-recognized. Consequently, the extent of treatment efficacy of an intervention can be different in different subgroups of patients. For example, patients with non-fatal MI and stroke are likely to respond better to a treatment if treated early than late. A genetic mapping and testing of virus in a virus-infected patient can tell whether this virus type will be resistant or susceptible to a given treatment. A breast cancer with estrogen receptor positive outcome can respond better to a treatment than a breast cancer estrogen receptor negative outcome. Herceptin responds better for metastatic breast cancer patients with an HER2 protein over-expression (Burstein, 2005). Thus, a trial, during randomization, can be enriched by a subgroup of patients who are likely to respond better to a given treatment than the rest of the patients of the trial. This can increase the success of the trial and can make the test more powerful for testing the treatment effect for the targeted subgroup. However, in this case, there is often concern that the overall treatment effect is driven mainly by the treatment effect in the subgroup and there is no treatment effect for the complement subgroup. If the review of the data of the complementary subgroup suggests that this is a possibility then the product label may reflect this. For example, the product label for atenolol (Tenormin) based on subgroup analyses included the statement, "Some subgroups (e.g., elderly patients with systolic blood pressure below 120 mm Hg) seemed less likely to benefit" (see Physicians' Desk Reference, 2002, Page 693). Such complimentary efficacy subgroup concern can also be addressed on using consistency-ensured multiple testing strategies discussed in Section 1.2.2.

### **1.2.7 Combination drug trials**

For fixed-combination drug products regulatory guidance requires that each component product, as mono-therapy, in the combination must be ef-

ficacious and safe and the combination product must demonstrate superior efficacy to each of its components in order to justify clinically meaningful benefit of the combination. This has led to trial designs with three arms, one arm for the combination and the other two for each of the components, provided that each component is an approved product for efficacy and safety. However, if the efficacy of any of the two components can not be assumed in a trial (e.g., for symptomatic treatments) or any of the two components is an unapproved product, then a 2-by-2 factorial design, which includes a placebo, is sometimes used. In this case, there is an extra burden to show that the unapproved component has clinically acceptable safety profile in addition to showing that it has a clinically meaningful efficacy.

The statistical test for these designs is the usual intersection-union test for showing that the combination is statistically significantly superior to each component, and if placebo is required in the trial, each component is also statistically significantly superior to placebo. Each of these tests is performed at the same significance level (Laska and Meisner, 1989; Sarkar et al., 1995).

The win criterion for efficacy for a drug combination trial can become complicated if two or more primary endpoints or three or more components are required in a trial to show the benefit of the combination over its approved components (Kwak et al., 2007). For example, for the case of two endpoints, the win criterion for efficacy of the combination may be to show superiority for one endpoint and at least non-inferiority for the other endpoint when comparing the combination to each component (superiority/non-inferiority procedures for problems of this kind are discussed in Section 4.6).

Sometimes multi-dose factorial designs are employed for the assessment of combination drugs for serving dual purpose, to provide confirmatory evidence that the combination is more effective than either component and to identify an effective and safe dose combination or a range of useful dose combinations (Hung, 2000).

### **1.2.8 Three-arm trials with an active control**

In many therapeutic areas, well-established standard treatments exist and for all new treatments seeking market authorization for the same indication there are then choices for the control treatment in clinical trials. If a standard treatment (active control) is selected this would lead to a two-arm study comparing the new treatment to the standard treatment for non-inferiority (using an appropriate non-inferiority margin) and (possibly) for superiority. There are well-known weaknesses connected with this design (see ICH E9) and therefore its use is recommended only for diseases where it is unethical to include placebo in the trial, and if sufficient historical data are available for the active control that can help in resolving the issues that these trials pose. These weaknesses come from the complexities of non-inferiority trial designs.

The goal for a standard non-inferiority trial is to infer indirectly about the efficacy of the new treatment on demonstrating that the new treatment is close



or similar (within a certain margin of non-inferiority) to the active control, which itself has previously been demonstrated to be effective by being superior to placebo. Thus, these trials have two comparisons, a direct comparison of the treatment against the active control and an indirect comparison against placebo which the trial is not able to include. However, the validity of the direct comparison, depends upon the validity of the indirect comparison and how much is known and how much can be assumed about the treatment effect of the active control in the setting of the current trial. Therefore, for assuring validity of this type of comparisons, the non-inferiority trials are required to have the following properties:

- Assay sensitivity of the trial. It is the ability of the trial to have shown a treatment difference of the active control in comparison to placebo of a specified size, if the trial had a third arm with placebo control. Without this property, the trial has the undesirable ability to conclude that the treatment is non-inferior to an ineffective drug.
- Constancy assumption. The trial is sufficiently similar to past historical studies with respect to all design and conduct features that can influence the estimation of the treatment effect of the active control, e.g., in regard to design features, patient population characteristics, important concomitant treatments, definition and ascertainments of study endpoints, dose of active control and background regimen, entry criteria, and analytic methods.
- Quality of the trial. This is in the interest of ruling out undesirable conduct features of the trial that would tend to minimize the difference between the treatment and the active control causing bias toward the null hypothesis. These include, for example, imprecise or poorly implemented entry criteria, poor compliance and the use of concomitant treatments whose effects may overlap with the treatment, in addition, inadequate measurement techniques, or errors in treatment assignments

Therefore, because of the above difficulties associated with non-inferiority trial designs, the use of three-arm trials with the new treatment, active control and placebo has gained much wider attention for diseases where it is ethical to include placebo in the trial. This type of trial has been called the “gold standard” design. The CHMP guideline “Choice of the non-inferiority margin” (2005) recommends this design to be used wherever possible. Pursuing all aims simultaneously in this type of trials raises the issue of multiple testing. There are several comparisons of interest:

- New treatment versus placebo for superiority.
- Active control versus placebo for superiority.
- New treatment versus active control for non-inferiority and (possibly) for superiority.

This approach is discussed in Koch and Röhmel (2004) and Röhmel and Pigeot (2009). See also Pigeot et al. (2003), Hauschke and Pigeot (2005) and discussion papers to this article. According to the closed testing methodology one could begin with an overall test for any differences between the groups which are followed by separate individual comparisons using the same significance level as for the overall test. Alternatively one could start with a test tailored to the many-to-one-situation (new treatment and active control versus placebo).

Pigeot et al. (2003) started with the comparison of the active control with placebo. Only after the active control has been shown to exhibit the expected superiority over placebo, Fieller's theorem (1954) is applied to the ratio of "net differences" (new treatment minus placebo and active control minus placebo). The limits of the resulting confidence interval allow a precise quantification of the fraction  $f$  of the effect of the active control that is preserved by the new treatment. A fraction  $f > 0$  is translated into "new treatment is superior to placebo", a fraction close to 1 means similar effects of the active control and new treatment, and a fraction larger than 1 confirms superiority of the new treatment over the active control.

If one is willing to accept a hierarchical structure between the hypotheses, one could also proceed as follows. Because the performance of the new treatment is of primary interest one could start immediately with the comparison between the new treatment and placebo. If the new treatment failed to demonstrate superiority over placebo, the remaining comparisons would lose much of their importance and would mainly serve for interpretation. Any superiority of the active control over placebo would not be surprising because it did so consistently in the past, and would reinforce the view that the new treatment is sufficiently effective. A failure of the active control, however, would raise doubts in the study itself (conduct, patient selection, sample size, or other design features).

The test for non-inferiority of the new treatment versus the active control is meaningful only in the case that one can successfully establish the superiority of the new treatment over placebo. Then it follows logically that two further null hypotheses of interest (the new treatment is inferior to the active control and the active control is no more effective than placebo) cannot both be true simultaneously. This means that both hypotheses can be tested using the full significance level, independently of each other. If the new treatment is non-inferior to the active control, a final test for superiority of the new treatment versus the active control can be carried out, again at the full level  $\alpha$ . This procedure controls the FWER in the strong sense. A similar problem was considered by D'Agostino and Heeren (1991) for which Dunnett and Tamhane (1992) showed how the sequence of the tests should be conducted so that the FWER is controlled. Active controlled trials that do not include placebo for ethical reasons, usually test for non-inferiority first, and if non-inferiority is established, then test for superiority next. This fixed-sequence testing has both clinical and statistical merit. The reverse fixed-sequence test-

ing, that is, test for superiority first and then for non-inferiority second, is generally not recommended for this type of trials. A reason for this is that the primary method for such a trial is usually the indirect demonstration of efficacy of the new treatment, as indicated above. In this framework, the non-inferiority hypothesis is primary and the superiority hypothesis is secondary. The sample size and design considerations of trials with this focus, and also the trial conduct rules and interpretation of results, are quite different than superiority trials. In addition, the determination of assay sensitivity and the non-inferiority margin can be quite complex and challenging.

### **1.2.9 Advanced multiplicity problems**

Multiple comparison problems discussed in the preceding sections dealt with a single source of multiplicity, e.g., two-arm trials with multiple endpoints or single-endpoint trials with multiple doses or subgroup analyses. However, multiplicity problems become more complex in trials with multiple sources of multiplicity, e.g., multiple treatment arms, multiple endpoints and tests for non-inferiority and superiority. The purpose of this section is to give examples of some such complex multiple comparison problems for some of which satisfactory solutions for regulatory applications have either not been worked out or available workable solutions are not sufficiently clear.

Example 1. A trial compares two doses of a new treatment to a control with respect to two primary efficacy endpoints and a quality of life (QoL) endpoint with all three endpoints considered equally important. The QoL endpoint is a composite (global) endpoint which includes four components that serve as secondary endpoints and can be tested only if the null hypothesis for the QoL endpoint is rejected.

Example 2. A trial compares four dose levels D1, D2, D3 and D4 of a new treatment to placebo on two primary endpoints for finding which dose levels are significantly better than placebo for both endpoints. In this comparison, the higher dose levels D3 and D4 are primary dose levels of equal importance and lower dose levels D1 and D2 are included to better understand the dose-response relationship in case the two higher dose levels are significant on both endpoints.

Example 3. A trial tests for treatment effects for the onset and durability endpoints at low, medium and high doses of a new treatment compared to placebo. In addition, the trial tests for certain loss in efficacy of effective doses at durability endpoints for putting patients on some sort of maintenance dose for continued meaningful treatment efficacy.

Example 4. A trial tests for treatment effects for multiple primary and secondary endpoints at low, medium and high doses of a new treatment compared to placebo with the restriction that tests for the secondary

endpoints for a specific dose can be carried out only when certain primary endpoints show meaningful treatment efficacy for that dose.

Example 5. A trial compares a new treatment to an active control on a primary and a secondary endpoint, and in this comparison, the trial performs non-inferiority and superiority tests in succession for the primary endpoint first. If non-inferiority or superiority are established for this primary endpoint, the trial then performs similar tests for the secondary endpoint.

Example 6. Instead of a single primary and a single secondary endpoint, as in Example 5, the trial has multiple primary and multiple secondary endpoints, and the non-inferiority or superiority test are carried out for certain secondary endpoints only if all primary endpoints show non-inferiority or superiority of the treatment against this control.

Example 7. A trial with a single primary endpoint includes three doses (high, medium and low) and an active control with non-inferiority and superiority tests for these doses. In this comparison, the trial sponsor faces two situations:

- Case 1. Dose effects can be assumed to be of non-decreasing order with increasing dose.
- Case 2. The assumption of a non-decreasing dose effect cannot be made, except perhaps in some special cases where only the low dose can be assumed to have efficacy not exceeding those of the medium and high doses.

A number of methods have been proposed to address the multiplicity problems in Examples 1 through 7. Hommel et al. (2007) proposed procedures based on the Bonferroni test that address multiplicity problems in Examples 1, 2 and 7. Further, procedures based on more powerful tests, e.g., Hochberg- or Dunnett-type tests, can be constructed using the general tree gatekeeping framework described in Section 5.5 (see also Dmitrienko, Tamhane and Liu, 2008). This framework allows testing of hierarchically ordered multiple families of hypotheses with logical restrictions. However, further research is needed to work out the details. Also note that solutions to the above problems can be far more challenging if trials were to also include interim analyses and allowed changes in certain design features based on interim results.

As an illustration, consider the multiple testing problem in Example 7. In Case 1, one may proceed with a fixed-sequence test, i.e., test for the high dose first for non-inferiority and then for superiority, and proceed to test similarly for the medium dose and then for the low dose if each time the preceding dose is found to be at least non-inferior to the active control. It has been argued that in this sequential testing, as two tests are performed for each dose, FWER can exceed  $\alpha$  if each test is performed at the same level  $\alpha$ . Therefore, for strong

control of FWER at level  $\alpha$ , appropriate multiplicity adjustment is warranted for this problem. Further, in Case 2, several approaches are possible. One approach would be to test first for non-inferiority simultaneously for all three doses, and then test for superiority only for those doses found to be non-inferior. A second approach would be to test for non-inferiority and then for superiority separately for each dose at adjusted  $\alpha$  levels, e.g., at  $\alpha/3$  for the Bonferroni method, or at levels that account for correlations among the test statistics. A third approach would be to define a family  $F_1$  of tests for the high and medium doses only and a family  $F_2$  of tests for the low dose only. Then assign  $\alpha_1$  to  $F_1$ , where  $\alpha_1 < \alpha$  (e.g.,  $\alpha_1 = 2\alpha/3$ ), and perform tests for non-inferiority and superiority for each dose in  $F_1$  as performed in the second approach above (e.g., spend  $\alpha/3$  for each dose test in  $F_1$ ). If a non-inferiority or superiority is established for a dose in  $F_1$  then  $\alpha$  used for that is basically saved which at least in part can then be carried forward to  $F_2$  on satisfying strong FWER control at level  $\alpha$ . As an example, if  $x$  doses in  $F_1$  were found to be superior to the control with the Bonferroni adjustment of  $\alpha/3$  for each dose, where  $x = 0, 1$  and  $2$ , then the low dose in  $F_1$  can be tested at level  $\alpha(x + 1)/3$ .

**Example 8.** A clinical trial requires multiplicity adjustments for comparing two doses of a new treatment to a placebo on two primary endpoints. These endpoints are able to characterize a clinically meaningful benefit of the treatment for each of these two doses with the condition that if only one of the endpoints is statistically significant for a dose then the other endpoint should exhibit at least a weak evidence of treatment benefit, that is, its  $p$ -value if greater than  $0.05$  is less than some pre-specified number  $\alpha^*$  (e.g.,  $\alpha^* = 0.25$ ). In addition, for this situation (i.e., when only one of the endpoints is statistically significant), a much stronger evidence of treatment benefit is required for the other endpoint. This problem can be solved by proper modifications of the consistency-ensured methods proposed in Section 1.2.2.

**Example 9.** A clinical trial uses a surrogate endpoint  $S$  for an accelerated approval and a clinically important endpoint  $T$  for a full approval. It uses  $\alpha_1$  for the accelerated approval and  $\alpha_2 = \alpha - \alpha_1$  for the full approval. In this case, a larger allocation for  $\alpha_1$  will increase the chance of an accelerated approval; however, it would impact negatively on the final approval under the underlying assumption that  $S$  is predictive of  $T$  only when the expected treatment effect for  $S$  exceeds certain limit. For example, an expected treatment effect, say of size  $x$ , for  $S$  would establish efficacy for  $S$  at a pre-specified level  $\alpha_1 = 0.01$ . However, such a value of  $x$  for  $S$  may not predict appropriate efficacy, say of size  $y$ , for the final endpoint  $T$  at  $\alpha_2 = 0.04$ . An appropriate procedure needs to be developed to take into account the strength of evidence in  $S$  (e.g., in terms of its  $p$ -value and other information) that will increase the chance

of a positive trial for the final endpoint  $T$  and will control the Type I error rate.

---

## 1.3 Reducing multiplicity in clinical trials

Testing many hypotheses in a trial can easily overload a trial with an excessive multiplicity burden making the trial unnecessarily complex, large and unrealistic from the cost and public health perspectives. Therefore, for meeting the objectives of a trial, consideration should be given in its design to reduce the burden of multiplicity efficiently. In this regard, several approaches are possible. A popular approach, discussed in Section 1.3.1, is to hierarchically order the families of hypotheses and also, as far as possible, the hypotheses within each family. This allows the use of new innovative statistical methods that adjust for multiplicity much more efficiently, in some cases with minimal or no adjustments at all. In addition, the use of dependence information among the test statistics can help in reducing the extent of multiplicity adjustments. This follows from the fact that the inflation of the FWER is generally largest when the test statistics are statistically independent, but relatively small when they are highly correlated. Another approach introduced in Section 1.3.2 is to combine several endpoints to define a single or a few composite endpoints.

### 1.3.1 Hierarchical testing

Families of null hypotheses are said to be hierarchically ordered or ranked if earlier families serve as gatekeepers in the sense that one tests hypotheses in a given family if the preceding gatekeepers have been successfully passed. The two commonly used hierarchical families of endpoints in a clinical trial are the family of primary endpoints and the family of secondary endpoints. These two families are hierarchically ordered with the property that rejections or non-rejections of null hypotheses of secondary endpoints depend on the outcomes of test results of primary endpoints. The individual endpoints within a family can also have hierarchical ordering, occurring naturally or by design. Hierarchical ordering of multiple endpoints and also of multiple comparisons can considerably reduce the multiplicity burden in controlling the FWER in a trial. The following examples further illustrate these points.

### Example: Gastric ulcer trial

Consider a gastric ulcer healing trial that compares a new treatment to a control using two endpoints that are healing of gastric ulcers (verified by endoscopy) and the symptom of gastric pain at the end of 16 weeks of treat-

ment. These two endpoints are naturally hierarchically ordered. That is, if the treatment heals all gastric ulcers in a patient then there is likely to be complete amelioration of gastric pain as well for that patient. In this case, one can use the fixed-sequence procedure (Section 2.6.3), i.e., test the healing endpoint first at the significance level of  $\alpha$ , and if this test is statistically significant then test for the symptom endpoint at the same significance level of  $\alpha$ . Both endpoints being statistically significant can strengthen the efficacy of the treatment of the trial.

### **Example: Congestive heart failure trial**

Consider a congestive heart failure (CHF) trial that accumulates events of mortality and hospitalizations over a two-year treatment period and has two primary endpoints. One endpoint is the composite of mortality and hospitalizations and the other is the mortality-only endpoint. These two endpoints can be ordered by design to test for the composite endpoint first thinking that it will have more events over the two-year treatment period for better power of the test for this endpoint. This situation, as in the above example, allows testing of the two endpoints sequentially at the same significance level.

However, a danger in sequential testing is that it can miss an important result. For example, if testing stops after the first test (i.e., the two-sided  $p$ -value for the composite endpoint is greater than 0.05) and the 2-sided  $p$ -value for the mortality endpoint happens to be small (e.g., less than 0.01) then this will incur a loss of a valuable trial result. For discussions following the results of a large clinical trial with outcomes for the two endpoints similar to this one, see

<http://www.fda.gov/ohrms/dockets/ac/03/briefing/3920B2.02.A-FDA-Coreg.pdf>

This danger can be reduced by using the fallback procedure discussed in Section 2.6.4. The fallback procedure has the flexibility of moving forward in the testing sequence even if some or all of the preceding tests in the testing sequence happen to be statistically non-significant. For example, for such a trial, one could test the composite endpoint at a pre-specified significance level of 0.035, and if the composite endpoint result is statistically significant at this level, then test the mortality-only endpoint at the full significance level of 0.05; otherwise, test the mortality endpoint at the reduced significance level of 0.015.

### **Example: A dose-finding hypertension trial**

For a rather complex example of hierarchical families of hypotheses involving multiple endpoints and multiple dose groups, consider a dose-finding study in patients with hypertension. Such a study was conducted to evaluate the effects of three doses, low, medium and high, of a treatment compared to

placebo where the effects were measured by reduction in systolic and diastolic blood pressure (SBP and DBP) measurements. For this study it was reasoned that

- SBP is more indicative of the true effect than DBP and hence it was placed at higher hierarchy.
- The medium and high doses were considered equally important, and potentially equally relevant, while the lower dose was considered less likely to exhibit significance.

Therefore, hierarchically ordered families of null hypotheses for this trial were set up as follows:

- $\{H_{11}, H_{12}\}$ : Null hypotheses for testing each of the high and medium doses compared to placebo with respect to SBP.
- $\{H_{21}, H_{22}\}$ : Null hypotheses for testing each of the high and medium doses compared to placebo with respect to DBP.
- $\{H_{31}\}$ : Null hypothesis for testing the low dose compared to placebo with respect to SBP.
- $\{H_{32}\}$ : Null hypothesis for testing the low dose compared to placebo with respect to DBP.

A parallel gatekeeping technique in Dmitrienko, Offen and Westfall (2003) provides a solution for this complex hierarchical testing problem (gatekeeping procedures are discussed in Chapter 5); an alternative solution based on the parametric fallback procedure is given in Huque and Alosh (2008) (see Section 4.3).

### 1.3.2 Composite endpoints

Another popular approach for reducing multiplicity, particularly for cardiovascular trials, is to combine several clinically relevant endpoints into a single endpoint, known as a composite endpoint. Chi (2005) cited three types of composite endpoints for clinical trials as primary endpoints:

- An index or a responder endpoint is constructed from multiple item scores, counts or from low yield endpoints. The HAM-D total score for depression trials and the ACR20 endpoint for rheumatoid arthritis trials are of this type. However, such a composite endpoint has to be clinically valid for the disease and the type of intervention under study. In addition, there should be sufficient experience from historical clinical studies to show that it has been reliable in detecting treatment effects, and has also been interpretable and clinically meaningful.



- A failure rate after a certain period of treatment or follow-up is computed on counting events that can occur from a predefined set of events. For example, in an organ transplant trial a failure during the six month of treatment can be a biopsy-proven acute rejection or graft loss or death.
- A few binary events are combined to form a composite event endpoint. Such a composite event endpoint generally arises in cardiovascular and chronic disease trials.

Chi (2005) made a distinction between the second and third types, but we see them as of the same type. Components of a composite endpoint are usually referred to as component or singleton endpoints. If  $E_1$  and  $E_2$  are binary events, then a composite event  $E$  is the union of these two events; i.e., it counts as an event if either  $E_1$  or  $E_2$  alone occurs, or both  $E_1$  and  $E_2$  jointly occur. For convenience in counting, one often counts only the first occurrence of the component events in a patient.

There are several motivations for using a composite event endpoint as a primary endpoint. It can reduce the size of the trial if the following conditions are met. The components in the composite increase the number of events in a non-overlapping manner; i.e., an event is not a direct consequence of the other. In addition, there is evidence for some homogeneity of treatment effects across the components of the composite or the components jointly enhance the treatment effect. A composite endpoint can also address broader aspects of a multifaceted disease. In this case, an isolated result in an endpoint may not be clinically meaningful. A composite endpoint can also allow changing the focus of the trial from discovering a large treatment effect to clinically meaningful small treatment effects that collectively can achieve a statistically significant benefit of the treatment. For further informative discussions of these motivations and relevant examples, see Moyé (2003).

Development of a composite endpoint at the time of trial design, however, requires special considerations and prior empirical evidence of its clinical value. Components of the composite are supposed to have some consistency with regard to their importance to patients, frequency of their occurrence, and the extent of treatment effects they are likely to produce. The use of a composite endpoint is generally discouraged for confirmatory clinical studies when large variations are expected to exist among its components with regard to these three aspects. The results of the component endpoints must also be fully disclosed and displayed along with the results of their composite for allowing a meaningful interpretation of the composite endpoint result.

Each component of a composite endpoint conveys a special meaning to the clinician. Therefore, statistical analysis of a composite endpoint generally includes some sort of analysis for its components. This raises multiplicity issues. However, statistical testing in this case usually follows a hierarchical structure; i.e., one tests for the composite endpoint first and then for its components. However, the multiplicity adjustment strategy for the components, depend-

ing on the objectives of the trial, can vary. It can vary from no adjustment for the analysis of components to adjustments to control the FWER in the strong sense. Therefore, multiple approaches for the analysis of components exist depending on the purpose of such analysis. Following are a few testing strategies that one can adopt in this regard.

Testing Strategy 1. In this case, the composite endpoint is tested at a pre-specified significance level  $\alpha$  and component endpoints are not statistically tested but their results are summarized for the interpretation of the composite endpoint using descriptive statistics that generally include point, interval estimates and unadjusted  $p$ -values for treatment effects. This approach has the disadvantage that if one of the components has a sufficiently small  $p$ -value, no specific claim of benefit can be made for this specific component endpoint, as there was no prospective plan for multiplicity adjustments for such a claim.

Testing Strategy 2. Use a sequential testing scheme. That is, if the composite endpoint result is statistically significant at the significance level of  $\alpha$ , then test for specified components in sequence at the same significance level  $\alpha$ . This method controls the FWER in the strong sense in testing all components. However, no claim can be made for a component even if it has a very small  $p$ -value if the sequence breaks, that is, when a test that precedes the testing of this particular component is not rejected. This deficiency can be avoided by using the fallback testing strategy, which allows testing of an important component of a composite, such as death, even when the null hypothesis for the composite endpoint is not rejected (see Section 2.6.4).

Testing Strategy 3. This is similar to Testing Strategy 2, except that instead of using a sequential testing scheme, one uses another method that controls the FWER in the strong sense. Examples include multiple testing procedures described in Section 2.6.

In resolving treatment effects, some components of a composite may have more persuasive abilities than others. The ones that have less persuasive abilities are often referred to as “soft components” of a composite endpoint. Showing a treatment benefit for the death endpoint (hard endpoint) can be more persuasive than showing a treatment benefit for the hospitalization endpoint. If the overall treatment effect is supported predominantly by the soft components then there should be assurance that the hard components are not influenced negatively. The following three approaches can address the issue of soft components of a composite endpoint.

The sub-composite approach. Create a “sub-composite” of those components of the “main composite” that are hard components. Then there are two null hypotheses to test one for the main composite and the other for the sub-composite. One can then test the null hypothesis for the main

composite at a smaller significance level  $\alpha_1$  and the sub-composite at a larger significance level  $\alpha_2$ , such that  $\alpha_1 + \alpha_2 = \alpha$ . For the above example, one may select  $\alpha_1 = 0.01$  and  $\alpha_2 = 0.04$ .

**Non-inferiority method.** Define a priori an acceptable margin for “sub-composite” and test for superiority for the main composite and test for non-inferiority for the “sub-composite”, both at the same significance level.

**The weighting method.** In this method, computation of the treatment effect estimate and the test statistic uses pre-determined  $\alpha$  “weights” given to the component endpoints with the sum of the weights being equal to one. For example, the weights selected for the harder endpoints could be three times larger than the weights assigned to softer endpoints. The advantage of this method is that it avoids adjusting for an extra sub-composite endpoint. However, these weights must be pre-defined and acceptable to clinicians.

There is a widespread interest for using composite endpoints as primary endpoints in a trial either for a single or multiple comparisons, especially for the cardiovascular type trials and for trials with patient reported outcomes that target claim of treatment benefits for a specific or multiple domains of a disease. This is because of the attractive nature of the composite endpoints in reducing multiplicity problems and in reducing the sample size of the trial. However, a cautionary point worth emphasizing in this regard is that trials with composite endpoints can be complicated when components of such endpoints are of widely differing importance to patients and occur with differing frequency, and are likely to vary markedly in the magnitude of treatment effects. Such situations can complicate interpreting the results of a trial. Therefore, when large such variations are likely to exist across components of composite endpoints, their use for controlled clinical trials can be problematic for making specific claims of treatment benefits.

---

## 1.4 Multiplicity concerns in special situations

### 1.4.1 Use of multiple subsets of patient data

It has been a common practice for many years that the primary statistical analysis concerns all randomized patients. The ideal situation that all patients are treated as intended by the randomization schedule and that all patients adhere to the study protocol is rare. Some irregularities usually occur. Patients may drop out of the study early, or patients may receive treatments different from the one assigned to them by the randomization schedule. Questions then arise as to how to deal with these irregularities. Because the true

measurements are not available, one must make assumptions. Often more than one plausible assumption is possible and sometimes a conservative view may require making more pessimistic assumptions. Therefore, several analyses must be performed on the same endpoint but with varying subsets of patients and/or varying subsets of data. As the reason of such “sensitivity analyses” is to demonstrate that the main study results are only marginally influenced by the observed irregularities, no adjustment for the Type I error rate is necessary. However, a different situation arises when following a statistical analysis for the whole study population subgroups are selected and subjected to separate statistical analyses. Multiplicity issues arising in this situation are discussed in detail in Section 1.2.6.

### **1.4.2 Use of multiple statistical methods**

Different statistical models or statistical methods are sometimes applied to the same data. In a parametric regression model one could, for example, include or exclude explanatory variables, or one could apply transformations to the data for achieving symmetry. It is also possible to change the analysis strategy from a parametric model to a nonparametric one or vice versa. Usually the primary analysis strategy is laid down in the Statistical Analysis Plan (SAP) that is finalized before the randomization schedule is opened. The analysis strategy as laid down in the SAP should contain sufficient details to avoid any ambiguity with the primary analysis. Technically one could also split the Type I error rate, if different analysis strategies are applied with control of the FWER. However, this has been rarely applied. In contrast, sometimes several analysis strategies are applied with the aim to demonstrate that the results are robust and are present regardless of the statistical model or method. Usually such procedures are not preplanned and as long as they do not indicate discrepant results they at least do no harm.

### **1.4.3 Analysis of covariance**

Unplanned analysis of covariance, similar to unplanned subgroup analysis (see [Section 1.2.6](#)), is prone to producing misleading results because of inherent multiplicity, bias and other issues. For this reason unplanned analysis of covariance is generally discouraged for drawing confirmatory evidence from a trial. Nonetheless, it is considered useful for exploratory purposes and for interpreting the results of a primary analysis. It can facilitate clarifying the degree to which the estimated treatment effects are due to the study treatment or by other factors that are associated with the response variable. It can also help in evaluating the consistency of treatment differences across subgroups. On the other hand, planned analysis of covariance is a useful method for randomized controlled clinical trials. If appropriately used, it can serve important purposes. For example, when comparing treatment groups, it can improve the power of the test under certain conditions through variance re-

duction resulting in smaller  $p$ -values and narrower confidence intervals. It can also adjust some of the random imbalances in baseline prognostic variables to “equivalence” in comparing treatment groups.

Planning and pre-specification of covariates and other aspects of this analysis, such as model specification, are considered essential. In addition, clear statistical as well as clinical justifications, stated in advance, are required for assuring that such an analysis will provide an unbiased estimate of the true difference between the treatments, and the covariate-adjusted treatment effect can be meaningfully interpreted for a clinical benefit. Covariates pre-selected for such an analysis need to be strongly correlated with the response variable. Covariates measured after randomization may be affected by the treatment – their inclusion into the analysis is likely to produce biased results. The CPMP (2003) document “Points to consider on adjustment for baseline covariates” provides a comprehensive list of recommendations to be followed when planning an analysis of covariance for a confirmatory trial. Also Pocock et al. (2002) and Senn (1994) provided useful comments and recommendations on this matter.

---

## 1.5 Multiplicity in the analysis of safety endpoints

This section includes some basic concepts and general framework for analysis of multiple safety endpoints. Safety evaluation of a treatment or an intervention intended for the treatment of a disease is an important objective of all clinical trials in general. For this reason, clinical trials collect adverse events (AEs) data for each patient of the trial along with other relevant information that help in clinical evaluation as well as quantitative analysis of these events. An AE refers to an “untoward medical event associated with the use of a drug in humans, whether or not it is considered as drug-related.” Some of these AEs can be serious. They are often referred to as serious adverse events (SAEs) or serious adverse drug experiences (SADEs). Examples of these are death, a life threatening event, an in-patient hospitalization, a persistent or significant disability, a congenital anomaly or a birth defect. Depending on the medical judgment, it may also include interventions such as medical or surgical procedures used for preventing serious safety events.

Statistical analysis of safety data includes analysis of particular events, analysis of event rates, evaluation of risk over time, exploration of possible subgroup differences, and the identification of risk factors associated with serious events. These analyses rely on those quantitative methods that can validly quantify risk and provide an adequate measure of uncertainty. The analysis approach usually taken is to analyze the safety data for each clinical study and then integrate the safety findings across multiple studies and other clinical experiences. In this effort, investigators group closely related events.

This is accomplished by using the so called the “dictionary” of preferred terms, such as the Medical Dictionary for Regulatory Activities (MedDRA), which was developed under the auspices of ICH. These analyses of safety data play an important role in determining the risk-benefit profile of a new treatment.

It is well recognized that analyses of safety data entail multidimensional problems involving many safety endpoints. A key point, however, is that not all these endpoints can be pre-specified. A number of them are unanticipated. This then calls for a different statistical inferential approach for different parts of the safety data. In this regard, it is worth recognizing that design and analysis approaches for safety assessments are in general different than those for efficacy assessments. For efficacy, primary and secondary efficacy endpoints are specified in advance. Trial sizes are usually determined with respect to primary endpoints, and serious efforts are made, in planning interim looks at data or in addressing multiplicity of tests for preserving the Type I error rate. On the other hand, in general, clinical trial practices have been not to specify safety hypotheses or the level of sensitivity in advance. However, there are exceptions. These are for situations when a particular safety concern related to a specific drug or drug class has risen or when there is a specific safety advantage being studied. For these cases, sometimes, trials for efficacy evaluation are large enough to address these safety issues, but every so often, when there is a significant safety concern, special large safety studies are conducted with pre-stated hypotheses and uncertainty specifications similar to efficacy trials. Considering these challenges and the ICH E9 ideas for safety data analysis, one usually classifies trials’ safety analysis strategies into three types:

Type 1. Analyses of adverse experiences associated with specific hypotheses that are formally tested in the clinical study and both Type I and Type II error concerns are properly addressed.

Type 2. Analyses of common adverse experiences which are usually treatment emergent signs and symptoms (TESS) not present at baseline, or if present at baseline, a greater severity seen when on treatment. Such AEs usually do not have pre-stated hypotheses, but their causality is assessed on comparing the incidence rates for patients treated with the test treatment versus those treated with the control. If the incidence rate of an AE is clearly greater with the test treatment than with the control, it can be attributed to the test treatment. For this purpose, clinical trials report  $p$ -values, point and/or interval estimates of risk differences or relative risks as descriptive statistics and for flagging purposes. However, methods differ when trials are short-term versus long-term studies. Long-term studies usually account for person-exposure times on treatment in the analysis with adjustments for censored observations. The inference approach may differ depending whether the intent is to report non-inferiority versus superiority.

Type 3. These refer to analyses of less common and rare spontaneous reports

of adverse events, some serious, often requiring attention of specialty area experts. Analysis of these usually require large data bases because these events may occur at very low rates, e.g., in the range of 1/100 to 1/1000.

Analysis methods for Type 2 AEs, for addressing multiplicity, are required to have certain special properties. For example, it should provide a proper balance between no adjustment and too much adjustment in the sense that one is willing to tolerate some wrongly flagged AEs provided their number is small relative to the total number of flagged AEs. In addition, it should address Type II error concerns and provide adjusted  $p$ -values for flagging purposes. In this regard, Mehrotra and Heyse (2004) recommended methods that provide control of the false discovery rate.

An approach that addresses Type II error concerns is to specify an error rate for failing to identify at least  $k$  (e.g.,  $k = 1$  or  $2$ ) unwanted safety events out of the total  $K$  analyzed; then for a specified  $\alpha$ , work out the size of the safety database needed for this purpose. If this size is limited then adjust the  $\alpha$  upward accordingly.

---

## 1.6 Concluding remarks

Clinical trials generally include multiple objectives, e.g., multiple efficacy and safety endpoints or comparisons of multiple doses of a new treatment to a control. The objectives are formulated to seek answers to a set of specified scientific questions. Whether answers to these questions can lead to claims of clinically meaningful benefits of the new treatment is determined by multiple “win” criteria that introduce multiplicity. Confirmatory trials prospectively plan statistical procedures that deal with these multiplicity issues in an efficient manner. Regulatory agencies in this regard usually ask for a well-developed and thought-out statistical analysis plan (SAP). If the multiplicity issues are complex, the SAP may include trial simulation results and sometimes mathematical proofs for assuring the validity and efficiency of the proposed statistical procedure.

A key point in this regard is that solutions to multiplicity problems generally require adhering to the principle of prospective planning, i.e., defining the multiplicity problems and working out their solutions in advance. Another key point is that, in the presence of multiplicity of tests, when specific claims of treatment benefits are intended for a new treatment, strong control of the FWER is almost always required.

In conclusion, as the subject matter of multiplicity for clinical trials is vast, and all cannot possibly be captured in an introductory chapter like this one even at the conceptual level, the scope of this chapter has been intentionally

limited to standard multiplicity topics that commonly arise in confirmatory clinical trials. Other multiplicity topics of complexity that arise, for example, for interim analyses and also for adaptive designs of confirmatory clinical trials, are also important multiplicity topics. These have been omitted here, but the reader may find some of them in other chapters of this book.



# Chapter 2

---

## *Multiple Testing Methodology*

**Alex Dmitrienko**

*Eli Lilly and Company*

**Frank Bretz**

*Novartis*

**Peter H. Westfall**

*Texas Tech University*

**James Troendle**

*National Institutes of Health*

**Brian L. Wiens**

*Alcon Laboratories*

**Ajit C. Tamhane**

*Northwestern University*

**Jason C. Hsu**

*Ohio State University*

---

### **2.1 Introduction**

Multiplicity issues are encountered in a variety of pharmaceutical applications with multiple objectives. In a pre-clinical setting, the objectives can correspond to multiple genetic markers. In clinical applications, the objectives can be defined in terms of multiple dose levels, endpoints or subgroup analyses. Most common sources of multiplicity in clinical trials are listed below:

- Multiple dose-control comparisons are commonly included in dose-finding studies to evaluate efficacy and safety properties of a treatment compared to a control.
- Multiplicity is often induced by multiple criteria for assessing the effi-

cacy profile of a treatment. These criteria are introduced to help the trial's sponsor better evaluate multiple dimensions of the treatment effect. Depending on the trial's objectives, the overall outcome may be declared positive if (i) one or more criteria are met or (ii) all criteria are met or (iii) some composite criterion is met.

- Another source of multiplicity in clinical trials is multiple secondary analyses, e.g., analysis of secondary endpoints or subgroup effects that are typically performed after the trial's primary objective is met.

This chapter provides an overview of key concepts and approaches in multiple testing methodology. For more information about the theory of multiple comparisons and a detailed review of popular multiple testing procedures and their properties, see Hochberg and Tamhane (1987), Westfall and Young (1993) and Hsu (1996).

The chapter is organized as follows. Sections 2.2 through 2.4 review fundamental concepts and principles that provide a foundation for the theory of multiple comparisons. Sections 2.5 through 2.8 describe commonly used multiple tests in pre-clinical and clinical studies. Lastly, Section 2.9 describes software packages that implement popular multiple testing procedures.

---

## 2.2 Error rate definitions

In order to choose an appropriate multiple testing method, it is critical to select the definition of correct decisions that reflect the objective of a clinical study or pre-clinical experiment. This section introduces an error rate definition commonly used in clinical trials (familywise error rate) as well as other definitions (generalized familywise error rate, false discovery rate and false discovery proportion) that have found applications in pre-clinical and clinical studies.

### 2.2.1 Comparisonwise and familywise error rates

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. As an example, consider a dose-finding study with  $m$  doses tested versus placebo. The primary endpoint is normally distributed with larger values indicating improvement. Let  $\mu_0$  be the mean improvement in the placebo arm and  $\mu_i$  be the mean improvement in the  $i$ th dose group,  $i = 1, \dots, m$ . The testing problem is formulated in terms of the difference in the mean responses: The hypothesis of treatment effect no greater than  $\delta$

$$H_i : \mu_i - \mu_0 \leq \delta$$

is tested versus a one-sided alternative

$$K_i : \mu_i - \mu_0 > \delta,$$

where  $\delta$  is a non-negative constant defining the clinically important difference. The Type I error rate for  $H_i$  is the probability of concluding that a clinically relevant treatment effect is present when the treatment difference is actually no greater than  $\delta$ .

If each of the  $m$  hypotheses is tested separately at a pre-specified significance level  $\alpha$ , e.g.,  $\alpha = 0.05$ , it can be shown that the proportion of incorrectly rejected hypotheses will not exceed  $\alpha$ . This is known as the control of the *comparisonwise error rate*. However, preserving the comparisonwise error rate is not considered an adequate approach to controlling the probability of incorrect decisions in a clinical trial setting. The hypotheses of interest are considered together as a family. Even a single Type I error in this family is assumed to lead to an incorrect decision. Accordingly, the overall Type I error rate is defined as the probability of rejecting at least one true hypothesis. The probability can be computed under the assumption that all  $m$  hypotheses are simultaneously true. This is known as the *weak control of the familywise error rate* (FWER). In the context of clinical trials with multiple endpoints, the weak FWER control can be interpreted as the probability of concluding an effect on at least one endpoint when there is no effect on any endpoint, i.e., the probability of concluding an ineffective treatment has an effect.

In general, the assumption that all hypotheses are true at the same time may be restrictive in many pharmaceutical applications and is not appropriate in the cases when the clinical trial sponsor is interested in making claims about specific outcomes. For example, in dose-finding clinical trials, the treatment difference is likely to vary across the dose levels and the trial's sponsor is generally interested in testing the drug effect at each particular dose and being able to claim that this effect is significant. To achieve this goal, one needs to preserve the probability of an incorrect decision (that is, the probability of erroneously finding a significant result) for each dose regardless of the size of the treatment effect in other dose groups. Using mathematical terminology, this requirement can be reformulated as the control of the probability of incorrectly rejecting any true hypothesis regardless of which and how many other hypotheses are true. In other words, if  $T$  is the index set of true null hypotheses, we require that

$$\sup \text{FWER} = \max_T \sup_{\{\mu_i(T)\}} P(\text{Reject at least one } H_i, i \in T) \leq \alpha,$$

where the supremum is taken over all  $\mu_i$  satisfying  $\mu_i - \mu_0 \leq \delta$  for  $i \in T$  and  $\mu_i - \mu_0 > \delta$  for  $i \notin T$ , and the maximum is taken over all index sets  $T$ . This approach to protecting the overall error rate is known as *strong control of the familywise error rate*. Strong control of the FWER for the primary objectives is mandated by regulators in all confirmatory clinical trials (CPMP, 2002). A

detailed description of multiple tests that protect the FWER in the strong sense is given in Sections 2.6–2.8.

It is worth noting that some multiple tests such as the unprotected and protected least significance difference (LSD) tests do not protect the FWER (Hochberg and Tamhane, 1987, Chapter 1). The former does not control the FWER even in the weak sense while the latter does not control it in the strong sense. These and similar tests will not be discussed in this chapter.

### 2.2.2 Generalized familywise error rate, false discovery rate and false discovery proportion

The definition of the FWER in the previous section is based on preserving the chances of rejecting at least one true null hypothesis for any number of hypotheses. This approach is reasonable when one deals with a few hypotheses but can become impractical in studies involving a large number of hypotheses, for example, in microarray experiments (Chapter 7). Specifically, as the number of hypotheses,  $m$ , increases, FWER-controlling multiple tests become conservative and fail to detect significant results unless the treatment effect is overwhelmingly positive.

The standard FWER definition can be extended by relaxing the requirement to protect the probability of at least one incorrect conclusion. This approach improves the power of multiple tests by increasing the probability of (correctly) rejecting false hypotheses. Romano, Shaikh and Wolf (2005) pointed out that, although one will need to pay a price for this in terms of an increased Type I error rate, “the price to pay can be small compared to the benefits to reap.”

#### Generalized familywise error rate

The *generalized familywise error rate* (gFWER) definition assumes that one can tolerate a certain fixed number  $k$  ( $1 \leq k < m$ ) of incorrect conclusions regardless of how many hypotheses are considered (Victor, 1982; Hommel and Hoffmann, 1987; Lehmann and Romano, 2005). In mathematical terms, the control of the generalized FWER is achieved if

$$\sup \text{gFWER}(k) = \max_T \sup_{\{\mu_i(T)\}} P(\text{Reject at least } k \text{ hypotheses } H_i, i \in T) \leq \alpha,$$

where  $T$  is the index set of at least  $k$  true null hypotheses. Note that the gFWER simplifies to the usual FWER when  $k = 1$ . Multiple testing procedures for controlling the gFWER are discussed in Chapter 7 in the context of problems arising in pharmacogenomic studies.

## False discovery rate and false discovery proportion

Two closely connected approaches to extend the FWER are known as the *false discovery rate* (FDR) (Benjamini and Hochberg, 1995) and the *false discovery proportion* (FDP) (Korn et al., 2004).

If the number of rejected hypotheses is positive, then the FDP is defined as

$$\text{FDP} = \left( \frac{\text{Number of rejected true null hypotheses}}{\text{Number of rejected hypotheses}} \right).$$

The FDP is defined as 0 if no hypotheses are rejected. The FDR is said to be controlled at the  $\gamma$  level if

$$\text{FDR} = E(\text{FDP}) \leq \gamma.$$

Note that control of the FDR at the  $\gamma$  level does not imply that the FDP is less than or equal to  $\gamma$  with high probability. To ensure this, one can choose an acceptable probability of exceedence,  $\alpha$ , and require that

$$P(\text{FDP} > \beta) \leq \alpha.$$

The interpretation is that of those hypotheses that are rejected, the proportion of false discoveries may exceed a specified fraction  $\beta$  with probability no larger than  $\alpha$ . Note that control of the FWER is equivalent to control of the FDP with  $\beta = 0$ . Control of the FDP makes sense in many nonconfirmatory settings like genetic or pre-clinical studies, where a certain proportion of errors is considered acceptable.

Control of the FDR at the  $\alpha$  level does not imply control of the FWER at the  $\alpha$  level, nor does any ( $\beta > 0$ ) control of the FDP at the  $\alpha$  level imply control of the FWER at the  $\alpha$  level. In fact, it is often possible to manipulate the design of a clinical trial so that any desired conclusion can be almost surely inferred without inflating the FDR (Finner and Roter, 2001). Thus, FDR or FDP controlling procedures are not suitable for confirmatory clinical trials.

### 2.2.3 Role of Type II errors

Similarly as in the case of the Type I error rate, the Type II error rate is not extended uniquely from the univariate case to the multiple-hypotheses case. Different possibilities exist to measure the success of a clinical trial in terms of power. A standard approach is to consider the probability of rejecting *at least one* false null hypothesis (disjunctive power), that is, to calculate

$$P(\text{Reject at least one } H_i, i \notin T),$$

where the probability is evaluated for a given set of parameter values:  $\mu_i > \mu_0 + \delta$  if  $i \notin T$  and  $\mu_i = \mu_0 + \delta$  if  $i \in T$  (Senn and Bretz, 2007). The use of disjunctive power is recommended in, for example, studies involving multiple comparisons with a control or in studies with multiple endpoints, where it

is sufficient to demonstrate the treatment's effect on at least one endpoint. Alternatively, one may be interested in calculating the probability of rejecting *all* false null hypotheses (conjunctive power)

$$P(\text{Reject all } H_i, i \notin T),$$

where probability is again evaluated at a given set of parameter values. One may argue that conjunctive power should be used in, for example, fixed drug combination studies or studies in which the treatment's effect must be established on two or more co-primary endpoints (see also Section 4.5).

Other power concepts exist and we refer to Maurer and Mellein (1987) and Westfall et al. (1999) for further details. General software implementations are not available, even for some of the simpler multiple comparison procedures and in most cases extensive simulations need to be performed. The practically relevant question about the appropriate power concept needs to be addressed on a case-by-case basis tailored to the study objectives, but see Hommel and Bretz (2008) for a balance between power and other considerations in multiple testing. It should be noted that adequately powering a clinical study is typically in the interest of the trial sponsor: It is the sponsor's choice to control the risk of failing to detect a truly significant drug effect.

## 2.3 Multiple testing principles

This section reviews key principles that provide a foundation for multiple tests described in this chapter. It begins with two general principles, known as the principles of *union-intersection testing* and *intersection-union testing*, that define the underlying testing problem. The section also introduces two methods for constructing multiple tests (*closure principle* and *partitioning principle*).

### 2.3.1 Union-intersection testing

Multiple testing problems in pharmaceutical applications are commonly formulated as union-intersection problems (Roy, 1953). Within the union-intersection framework, one rejects the global hypothesis of no effect if there is evidence of a positive effect with respect to at least one individual objective. To provide a mathematical definition, let  $H_1, \dots, H_m$  denote the hypotheses corresponding to the multiple objectives. The hypotheses are tested against the alternative hypotheses  $K_1, \dots, K_m$ . The global hypothesis  $H_I$ , defined as the intersection of the hypotheses, is tested versus the union of the alternative

hypotheses ( $K_U$ ):

$$H_I : \bigcap_{i=1}^m H_i \text{ versus } K_U : \bigcup_{i=1}^m K_i.$$

In the context of union-intersection testing, carrying out the individual tests at an unadjusted  $\alpha$  level leads to an inflated probability of rejecting  $H_I$  and can compromise the validity of statistical inferences. To address this problem, a multiplicity adjustment method needs to be utilized to control the appropriately defined probability of a Type I error.

### 2.3.2 Intersection-union testing

A different class of multiple testing problems requires a different approach called the intersection-union testing approach. Intersection-union testing arises naturally in studies when a significant outcome with respect to two or more objectives is required in order to declare the study successful. For example, new therapies for the treatment of Alzheimer's disease are required to demonstrate their effects on both cognition and global clinical scores.

In other words, the intersection-union method involves testing the union of the hypotheses ( $H_U$ ) against the intersection of the alternative hypotheses ( $K_I$ ):

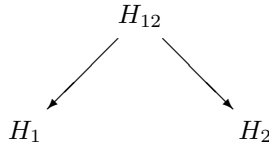
$$H_U : \bigcup_{i=1}^m H_i \text{ versus } K_I : \bigcap_{i=1}^m K_i.$$

When the global hypothesis  $H_U$  is rejected, one concludes that all  $K_i$ 's are true, i.e., there is evidence of a positive effect with respect to all of the  $m$  objectives.

An interesting feature of intersection-union tests is that no multiplicity adjustment is necessary to control the size of a test but the individual hypotheses cannot be tested at levels higher than the nominal significance level either (Berger, 1982). Note that intersection-union tests are sometimes biased in the sense that their power function can drop below their size (Type I error rate) in the alternative space. For a detailed discussion of intersection-union tests in the analysis of multiple endpoints, see Section 4.5.

### 2.3.3 Closure principle

The closure principle proposed by Marcus, Peritz and Gabriel (1976) plays a key role in the theory of multiple testing and provides a foundation for virtually all multiple testing methods arising in pharmaceutical applications. This principle has been used to construct a variety of stepwise testing procedures.



**FIGURE 2.1:** Closed testing procedure in the dose-finding clinical trial example with two null hypotheses.

### Motivating example

To introduce the closure principle and demonstrate how it can be used to derive powerful multiple tests, consider the dose-finding trial example described in Section 2.2.1. Assume that two doses are compared to a placebo ( $m = 2$ ) and the clinically important difference  $\delta$  is equal to 0. The two associated hypotheses are given by  $H_1 : \mu_1 \leq \mu_0$  and  $H_2 : \mu_2 \leq \mu_0$ . Let  $p_1$  and  $p_2$  denote the  $p$ -values for testing  $H_1$  and  $H_2$ .

To construct a closed testing procedure for  $H_1$  and  $H_2$ , we first need to define the *closed family of hypotheses* by considering all possible intersections of the two hypotheses. This family contains the following three intersection hypotheses:

$$H_1, H_2 \text{ and } H_{12} = H_1 \cap H_2.$$

The closure principle states that an FWER-controlling testing procedure can be constructed by testing each hypothesis in the closed family using a suitable *local*  $\alpha$ -level test. This procedure rejects a hypothesis if all intersection hypotheses containing this hypothesis are rejected by the associated local tests. The decision rule is depicted in Figure 2.1. To conclude that  $\mu_1 > \mu_0$ , the two intersection hypotheses containing  $H_1$ , i.e.,  $H_1$  and  $H_{12}$ , need to be rejected. Likewise, to show that  $\mu_2 > \mu_0$ ,  $H_2$  and  $H_{12}$  need to be rejected.

As an illustration, we will construct a closed testing procedure based on the Bonferroni test. Let  $\alpha$  denote the significance level, e.g.,  $\alpha = 0.05$ . The local tests reject  $H_1$  and  $H_2$  if  $p_1 \leq \alpha$  and  $p_2 \leq \alpha$ , respectively. Further, the Bonferroni test is carried out to test  $H_{12}$ , i.e., the hypothesis is rejected by its local test if  $p_1 \leq \alpha/2$  or  $p_2 \leq \alpha/2$ . When the decision rules are combined, it is easy to show that the resulting procedure has the following form:

- The procedure begins with the more significant  $p$ -value and rejects the corresponding hypothesis if the  $p$ -value is no greater than  $\alpha/2$ .
- If the first hypothesis is rejected, the other hypothesis is tested at an  $\alpha$  level. Otherwise, the other hypothesis is retained without testing.

Since the second test is carried out at the  $\alpha$  level, the closed testing procedure rejects as many (and sometimes more) hypotheses as the Bonferroni test from which it was derived. The power gain is due to the application of the



closure principle. As a side note, this closed testing procedure is actually the stepwise testing procedure proposed by Holm (1979). For more information about the Bonferroni and Holm procedures, see [Section 2.6](#).

## General case

In the general case of testing  $m$  hypotheses, the process of constructing a closed testing procedure goes through the following steps:

- Define the closed family of hypotheses. For each non-empty index set  $I \subseteq \{1, \dots, m\}$ , consider an intersection hypothesis defined as

$$H_I = \bigcap_{i \in I} H_i.$$

- Establish implication relationships. An intersection hypothesis that contains another intersection hypothesis is said to imply it, i.e.,  $H_I$  implies  $H_J$  if  $J \subset I$ .
- Define local  $\alpha$ -level tests for individual intersection hypotheses. Let  $p_I$  denote the  $p$ -value produced by the associated local test and reject  $H_I$  iff  $p_J \leq \alpha$  for all  $J \subseteq I$ . In particular, reject  $H_i$  if and only if (iff) all intersection hypotheses containing  $H_i$  are rejected by their local tests. In other words, reject  $H_i$  if and only if  $p_I \leq \alpha$  for all index sets  $I$  that include  $i$ .

Marcus et al. (1976) showed that this closed testing procedure for the hypotheses  $H_1, \dots, H_m$  controls the FWER in the strong sense at the  $\alpha$  level. To see how closed testing procedures achieve strong FWER control, it is instructive to revisit the FWER definition given in Section 2.2. By considering all possible combinations of hypotheses and defining an  $\alpha$ -level test for each intersection, we ensure that the resulting procedure protects the Type I error rate for any configuration of true hypotheses. This immediately implies that the FWER is controlled in the strong sense.

The closed testing algorithm is generally computationally intensive since approximately  $2^m$  individual tests need to be carried out to test  $m$  hypotheses. Because of this, shortcut versions of closed testing procedures have attracted attention in the multiple testing literature. Shortcut procedures have a stepwise form and reduce the number of computational steps from order- $2^m$  to order- $m$  or order- $m^2$ . In addition, as will be explained in Section 2.6, stepwise procedures provide useful insights into the process of performing multiplicity adjustments and are easy to communicate to non-statisticians. For more information on stepwise closed testing procedures, see Grechanovsky and Hochberg (1999), Westfall et al. (2001), Dmitrienko et al. (2006b), Hommel, Bretz and Maurer (2007) and Bretz et al. (2009b).

### 2.3.4 Properties of closed testing procedures

This section briefly describes important properties of closed testing procedures that will be referenced later in this chapter and other chapters.

#### Monotone procedures

A monotone procedure rejects a hypothesis whenever it rejects another hypothesis with a larger  $p$ -value. For example, if  $p_i < p_j$  then the rejection of  $H_j$  automatically implies the rejection of  $H_i$ . Monotonicity helps to avoid logical inconsistencies; as such it is an essential requirement for multiple testing procedures. When a procedure does not have this property, monotonicity needs to be enforced by updating adjusted  $p$ -values. The Shaffer procedure introduced in Section 2.6.2 serves as an example of a procedure that requires monotonicity to be enforced. For a review of other monotonicity considerations, see Hommel and Bretz (2008).

#### Consonant procedures

A closed testing procedure is termed consonant (Gabriel, 1969) if the rejection of an intersection hypothesis  $H_I$  with  $I \subseteq \{1, \dots, m\}$  and  $|I| > 1$  always leads to the rejection of at least one  $H_J$  implied by  $H_I$ , i.e.,  $H_J$  with  $J \subset I$ . While consonance is generally desirable, nonconsonant procedures can be of practical importance. The Hommel procedure defined in Section 2.6.8 is an example of a nonconsonant closed testing procedure. It is possible for this procedure to reject the global null hypothesis  $H_I$ ,  $I = \{1, \dots, m\}$ , without rejecting any other intersection hypotheses.

#### $\alpha$ -exhaustive procedures

An  $\alpha$ -exhaustive procedure is a closed testing procedure based on intersection hypothesis tests the size of which is exactly  $\alpha$  (Grechanovsky and Hochberg, 1999). In other words,  $P(\text{Reject } H_I) = \alpha$  for any intersection hypothesis  $H_I$ ,  $I \subseteq \{1, \dots, m\}$ . If a procedure is not  $\alpha$ -exhaustive, one can construct a uniformly more powerful procedure by setting the size of all intersection hypothesis tests at  $\alpha$ . It is worth noting that some popular multiple testing procedures, for example, the fallback and Hochberg procedures described in Sections 2.6.4 and 2.6.9, respectively, are not  $\alpha$ -exhaustive. These procedures are used in pharmaceutical applications due to other desirable properties such as computational simplicity.

### 2.3.5 Partitioning principle

The partitioning principle was introduced by Stefansson, Kim and Hsu (1988) and Finner and Strassburger (2002). The advantage of using this principle is two-fold:

- It can be used to construct procedures that are more powerful than procedures derived using the closed testing principle.
- Partitioning procedures are easy to invert in order to set up simultaneous confidence sets for parameters of interest (these sets are constructed by inverting partitioning tests as explained in Section 2.4.2).

### Motivating example

To illustrate the process of carrying out partitioning tests, consider the clinical trial example with two doses and a placebo from Section 2.3.3. The first step involves partitioning the union of the hypotheses

$$H_1 : \mu_1 \leq \mu_0, \quad H_2 : \mu_2 \leq \mu_0.$$

into three mutually exclusive hypotheses:

$$\begin{aligned} H_1^* : \mu_1 \leq \mu_0 & \quad \text{and} \quad \mu_2 \leq \mu_0, \\ H_2^* : \mu_1 \leq \mu_0 & \quad \text{and} \quad \mu_2 > \mu_0, \\ H_3^* : \mu_1 > \mu_0 & \quad \text{and} \quad \mu_2 \leq \mu_0. \end{aligned}$$

Since the three hypotheses are disjoint, each one of them can be tested at level  $\alpha$  without compromising the FWER control. The final decision rule is constructed by considering all possible outcomes for the three mutually exclusive hypotheses. For example,

- If  $H_1^*$  is rejected, we conclude that  $\mu_1 > \mu_0$  or  $\mu_2 > \mu_0$ .
- If  $H_1^*$  and  $H_2^*$  are rejected, we conclude that  $\mu_1 > \mu_0$  and, similarly, rejecting  $H_1^*$  and  $H_3^*$  implies that  $\mu_2 > \mu_0$ .
- If  $H_1^*$ ,  $H_2^*$  and  $H_3^*$  are all rejected, the conclusion is that  $\mu_1 > \mu_0$  and  $\mu_2 > \mu_0$ .

This test appears conceptually similar to the closed test described in Section 2.3.3. However, unlike the closure, the partitioning principle does not deal with the hypotheses in the closed family (i.e.,  $H_1$ ,  $H_2$  and  $H_1 \cap H_2$ ) but rather with mutually exclusive hypotheses that partition the union of  $H_1$  and  $H_2$ .

## General case

To briefly describe a general version of the partitioning principle, let  $\theta$  be the  $k$ -dimensional parameter of interest in a pre-clinical experiment or a clinical trial,  $k \geq 1$ . Suppose  $m$  hypotheses are considered and assume that  $H_i$  states that  $\theta \in \Theta_i$ , where  $\Theta_i$  is a subset of the  $k$ -dimensional space,  $i = 1, \dots, m$ . For example, in the dose-finding example discussed above,  $\theta$  is a three-dimensional vector of true treatment means,  $\theta = (\mu_0, \mu_1, \mu_2)$ , and  $H_1$  and  $H_2$  are formulated in terms of

$$\Theta_1 = \{(\mu_0, \mu_1, \mu_2) : \mu_1 \leq \mu_0\}, \quad \Theta_2 = \{(\mu_0, \mu_1, \mu_2) : \mu_2 \leq \mu_0\}.$$

Given  $\Theta_1, \dots, \Theta_m$ , partition the union of the  $m$  subsets into disjoint subsets  $\Theta_I^*$ ,  $I \subseteq \{1, \dots, m\}$ . Each subset can be interpreted as a part of the  $k$ -dimensional space in which the hypotheses  $H_i$ ,  $i \in I$ , are true and the remaining hypotheses are false. The next step is to define hypotheses corresponding to the constructed subsets,

$$H_I^* : \theta \in \Theta_I^*,$$

and test them at the  $\alpha$  level. Since these hypotheses are mutually exclusive, at most one of them is true. Thus, even though no multiplicity adjustment is made, the resulting multiple test controls the FWER at the  $\alpha$  level.

For more information about the partitioning principle and its applications to multiplicity problems in pre-clinical and clinical studies, see Hsu and Berger (1999), Xu and Hsu (2007), Strassburger, Bretz and Finner (2007) and Strassburger and Bretz (2008).

---

## 2.4 Adjusted significance levels, $p$ -values and confidence intervals

Multiple inferences are performed by adjusting decision rules for individual hypotheses. This can be accomplished by computing multiplicity-adjusted significance levels, multiplicity-adjusted  $p$ -values or simultaneous confidence intervals. To avoid inflation of the overall Type I error rate in multiple testing problems, significance levels for individual hypotheses are adjusted downward or  $p$ -values are adjusted upward. Similarly, wider confidence intervals for parameters of interest need to be chosen to keep the overall coverage probability at a pre-determined level.

### 2.4.1 Adjusted significance levels and $p$ -values

In most simple cases, a multiplicity adjustment can be performed by computing a reduced significance level for each individual hypothesis. For example,

in the problem of testing the hypotheses  $H_1, \dots, H_m$ , a multiple test can be carried out by comparing the  $p$ -value associated with  $H_i$  to a significance level,  $\alpha_i$ , which is lower than the nominal  $\alpha$  level. The  $\alpha_i$ 's are selected to maintain the FWER at the  $\alpha$  level.

In general, adjusted significance levels are used less frequently than adjusted  $p$ -values, mainly because adjusted significance levels depend on the  $\alpha$  level. However, there are cases when the use of adjusted significance levels simplifies multiplicity adjustments. Consider, for example, a meta analysis that combines several multinational studies. If different multiplicity adjustment strategies are required by different regulatory agencies, the meta analysis may be easier to implement using raw  $p$ -values with appropriately adjusted significance levels.

Unlike adjusted significance levels, adjusted  $p$ -values capture the degree of multiplicity adjustment without reference to the pre-specified error rate and thus one can choose different  $\alpha$  levels for different sets of hypotheses. For example, a clinical trial sponsor can pre-specify the 0.05 significance level for hypotheses corresponding to the primary objectives and a higher level (e.g.,  $\alpha = 0.1$ ) for secondary hypotheses. Another advantage of adjusted  $p$ -values is that they incorporate the structure of the underlying decision rule which can be quite complex. Considerations of this type become important, for example, in the context of gatekeeping procedures described in Chapter 5.

A general definition of an adjusted  $p$ -value is given in Westfall and Young (1993): The adjusted  $p$ -value for a hypothesis is the smallest significance level at which one would reject the hypothesis using the given multiple testing procedure. This definition can be illustrated by applying it to closed testing procedures. As was explained in Section 2.3.3, a closed testing procedure rejects a hypothesis, for example,  $H_i$ , if all intersection hypotheses containing  $H_i$  are rejected. If  $p_I$ ,  $I \subseteq \{1, \dots, m\}$ , denotes the  $p$ -value for testing the intersection hypothesis  $H_I$ , the adjusted  $p$ -value for  $H_i$  is the largest  $p$ -value associated with the index sets including  $i$ :

$$\tilde{p}_i = \max_{I: i \in I} p_I.$$

The hypothesis  $H_i$  is rejected if the adjusted  $p$ -value does not exceed the pre-specified  $\alpha$  level, i.e.,  $\tilde{p}_i \leq \alpha$ . This general approach will be utilized in Sections 2.6–2.8 to derive adjusted  $p$ -values for multiple testing procedures commonly used in pharmaceutical applications (all of which can be formulated as closed testing procedures).

## 2.4.2 Simultaneous confidence intervals

Lehmann (1986, page 90) described the following general method for constructing a confidence set from a significance test. Let  $\theta$  denote the parameter of interest. For each parameter value  $\theta_0$ , test the hypothesis  $H : \theta = \theta_0$  using an  $\alpha$ -level test and then consider the set of all parameter values  $\theta_0$  for which

$H : \theta = \theta_0$  is retained. The set is, in fact, a  $100(1 - \alpha)\%$  confidence set for the true value of  $\theta$ . This method is essentially based on partitioning the parameter space into subsets consisting of a single parameter point each.

In the context of multiple hypothesis testing the partitioning principle described in Section 2.3.5 provides a natural extension of this general method to derive simultaneous confidence intervals that are compatible with a given multiple testing procedure (Hayter and Hsu, 1994; Finner and Strassburger, 2002).

Applying the partitioning principle, the parameter space is partitioned into small disjoint subhypotheses, each of which is tested with an appropriate test. The union of all non-rejected hypotheses then yields a confidence set  $C$  for the parameter vector of interest (see Finner and Strassburger, 2002, for a formal description). Note that the finest possible partition is given by a pointwise partition such that each point of the parameter space represents an element of the partition. Most of the classical (simultaneous) confidence intervals can be derived by using the finest partition and an appropriate family of one- or two-sided tests. However, this is not true in general. Note that a confidence set  $C$  can always be used to construct simultaneous confidence intervals by simply projecting  $C$  on the coordinate axes. Compatibility can be ensured by enforcing mild conditions on the partition and the test family (Strassburger, Bretz and Hochberg, 2004). In the following sections we will define simultaneous confidence intervals for popular multiple testing procedures. We will see that simultaneous confidence intervals are easily obtained for single-step procedures, but are often difficult to derive for stepwise procedures.

---

## 2.5 Common multiple testing procedures

This section provides background information and sets the stage for the next three sections (Sections 2.6–2.8) which review popular multiple testing procedures in pharmaceutical applications. We will begin by introducing several possible classification schemes based on the testing sequence, distributional assumptions and control of the Type I error rate.

### 2.5.1 Classification of multiple testing procedures

#### Single-step and stepwise procedures

Two important types of multiple testing procedures considered in Sections 2.6–2.8 are *single-step* and *stepwise* procedures described below.

Single-step procedures are multiple testing procedures for which the decision to reject any hypothesis does not depend on the decision to reject any other hypothesis. In other words, the order in which the hypotheses are

tested is not important and one can think of the multiple inferences as being performed simultaneously in a single step. The Bonferroni procedure (Section 2.6.1) and Dunnett procedure (Section 2.7.1) are examples of single-step procedures.

Unlike single-step procedures, stepwise procedures are carried out in a sequential manner. Some hypotheses are not tested explicitly and may be retained or rejected by implication. Stepwise procedures provide an attractive alternative to single-step procedures because they can reject more hypotheses without inflating the overall error rate.

The stepwise testing approach can be implemented via *step-down* or *step-up* procedures:

- A step-down procedure starts with the most significant  $p$ -value and continues in a sequentially rejective fashion until a certain hypothesis is retained or all hypotheses are rejected. If a hypothesis is retained, testing stops and the remaining hypotheses are retained by implication. The Holm procedure is an example of a step-down testing procedure.
- Step-up procedures approach the hypothesis testing problem from the opposite direction and carry out individual tests from the least significant one to the most significant one. The final decision rule is reversed compared to step-down procedures; i.e., once a step-up procedure rejects a hypothesis, it rejects the rest of the hypotheses by implication. The Hochberg procedure is an example of a step-up testing procedure.

The Holm and Hochberg procedures mentioned above are defined in Sections 2.6.2 and 2.6.9, respectively.

## Distributional assumptions

Another useful approach to the classification of multiple testing procedures is based on the assumptions they make about the joint distribution of the test statistics. This approach leads to the following classification scheme:

- Procedures that don't make any assumptions about the joint distribution of the test statistics. These procedures rely on univariate  $p$ -values and thus tend to have a rather straightforward form. They are referred to as  *$p$ -value based procedures* or nonparametric procedures. Examples include many popular procedures such as the Bonferroni and Holm procedures. These and similar procedures are discussed in Section 2.6.
- Procedures that make specific distributional assumptions, for example, that the test statistics follow a multivariate normal or  $t$ -distribution. To contrast this approach with nonparametric procedures based on univariate  $p$ -values, they are termed *parametric procedures*. Examples include the Dunnett and related procedures introduced in Section 2.7.

- Procedures that do not make specific assumptions and attempt to approximate the true joint distribution of the test statistics. The approximation relies on resampling-based methods (bootstrap or permutation methods) and thus procedures in this class are often referred to as *resampling-based procedures*. The resampling-based approach is described in Section 2.8.

It is important to point out that  $p$ -value-based procedures tend to perform poorly, compared to parametric and resampling-based procedures, when the testing problem involves a large number of hypotheses or the test statistics are strongly correlated. This is due to the fact that procedures that do not account for the correlation among test statistics become conservative in these cases.

## Control of the Type I error rate

The multiple testing procedures described in Sections 2.6–2.8 focus on the strong control of the FWER. Procedures that control alternative error rate definitions, e.g., the generalized FWER, are discussed in Chapter 7.

### 2.5.2 Notation

The following notation will be used in this section. As before,  $H_1, \dots, H_m$  denote the hypotheses of interest. We will assume throughout Sections 2.6–2.8 that the  $m$  hypotheses are tested under the free combination condition; i.e., no logical dependencies exist among the hypotheses. The only two exceptions are the problems considered in Section 2.6.2 (Shaffer procedure) and Section 2.7.3 (extended Shaffer-Royen procedure).

When the hypotheses are not equally important, the weights,  $w_1, \dots, w_m$ , are introduced to quantify their importance (each weight is between 0 and 1 and the weights add up to 1). Weighted hypotheses are encountered, for example, in dose-finding trials. The trial's sponsor can assign weights to the dose-placebo comparisons according to the expected effect size at each dose level to improve the overall power of the multiple test.

The test statistics associated with the hypotheses are denoted by  $t_1, \dots, t_m$ . Let  $p_i$  be the  $p$ -value computed from the null distribution of  $t_i$ ,  $i = 1, \dots, m$ . These  $p$ -values are frequently called *raw  $p$ -values* to distinguish them from multiplicity adjusted  $p$ -values. The ordered  $p$ -values are denoted by  $p_{(1)} < \dots < p_{(m)}$  and the associated hypotheses are denoted by  $H_{(1)}, \dots, H_{(m)}$ .

### 2.5.3 Dose-finding trial example

To illustrate the use of multiple testing procedures in pharmaceutical applications, we will use the following example. Consider a dose-finding trial in



**TABLE 2.1:** Summary of the mean increase in HDL cholesterol (mg/dl) in the dose-finding trial under three scenarios (mean difference, standard error, lower limit of the one-sided 97.5% confidence limit, two-sample  $t$  statistic and raw one-sided  $p$ -value). The asterisk identifies the  $p$ -values that are significant at the 0.025 level.

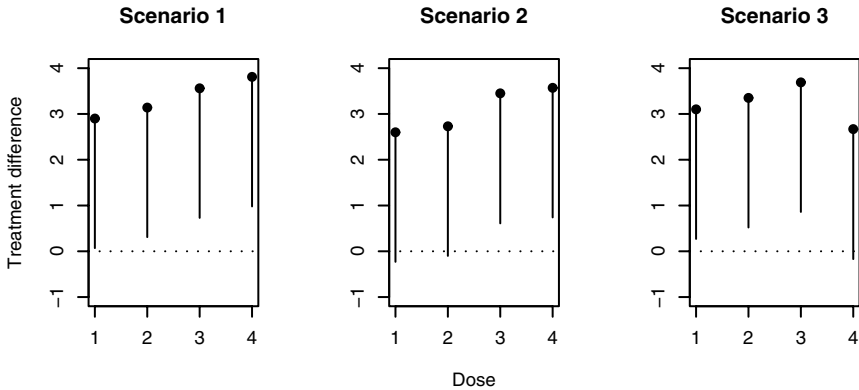
Test	Mean	Standard error	Lower confidence limit	$t$ statistic	$P$ -value
Scenario 1					
D1-Placebo	2.90	1.44	0.07	2.01	0.0228*
D2-Placebo	3.14	1.44	0.31	2.17	0.0152*
D3-Placebo	3.56	1.44	0.73	2.46	0.0071*
D4-Placebo	3.81	1.44	0.98	2.64	0.0043*
Scenario 2					
D1-Placebo	2.60	1.45	-0.23	1.80	0.0364
D2-Placebo	2.73	1.45	-0.10	1.89	0.0297
D3-Placebo	3.45	1.45	0.61	2.38	0.0088*
D4-Placebo	3.57	1.45	0.74	2.47	0.0070*
Scenario 3					
D1-Placebo	3.10	1.45	0.27	2.15	0.0162*
D2-Placebo	3.35	1.45	0.52	2.32	0.0105*
D3-Placebo	3.69	1.45	0.86	2.55	0.0055*
D4-Placebo	2.67	1.45	-0.17	1.85	0.0329

patients with dyslipidemia. The trial will be conducted to compare the effect of four doses of the drug, 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 77 patients.

Table 2.1 displays the mean treatment effects of the four doses compared to placebo, associated standard errors, lower limits of one-sided 97.5% confidence intervals,  $t$  statistics based on the two-sample  $t$  test with a pooled variance computed from all treatment groups and raw one-sided  $p$ -values. The table includes three scenarios that represent three different dose-response relationships in this trial. These scenarios will be used to evaluate the performance of the multiple testing procedures described in Sections 2.6–2.8.

The mean treatment differences with one-sided 97.5% confidence intervals in the three scenarios are plotted in [Figure 2.2](#). Key features of the three dose-response functions are summarized below:

- Scenario 1. The dose-response function increases over the dose range and the drug effect is present in all dose groups (all doses are superior to placebo at 0.025).
- Scenario 2. The dose-response function increases over the dose range



**FIGURE 2.2:** Mean treatment differences in HDL cholesterol change from baseline to 12 weeks with one-sided 97.5% confidence intervals in the dose-finding trial example.

but the drug effect is present only at the higher doses (D3 and D4 are superior to placebo but D1 and D2 are not).

- Scenario 3. The dose-response function is generally flat at the lower three doses with a drop at the highest dose due to tolerability problems (D1, D2 and D3 are superior to placebo but D4 does not separate from placebo).

It should be emphasized that we use the numerical example from [Table 2.1](#) mainly to illustrate the multiple testing procedures described in Sections 2.6–2.8. Chapter 3 describes alternative analysis strategies based on multiple testing, modeling dose-response functions or a combination of multiple testing and modeling, which are usually more appropriate in the context of dose-finding studies.

## 2.6 Multiple testing procedures based on univariate $p$ -values

### 2.6.1 Bonferroni procedure

The Bonferroni procedure is a widely used single-step procedure commonly attributed to Sir Ronald Fisher. In the problem of testing  $m$  equally weighted

hypotheses,  $H_1, \dots, H_m$ , the Bonferroni procedure rejects  $H_i$  if  $p_i \leq \alpha/m$ . Due to the (first-order) Bonferroni inequality, this procedure controls the FWER for any joint distribution of the raw  $p$ -values.<sup>1</sup>

As an illustration, consider Scenario 1 of the dose finding trial example given in Section 2.5.3. The Bonferroni-adjusted significance level is  $\alpha/4 = 0.00625$  and thus the D4-Placebo test is significant at this level whereas the other three tests are not.

The Bonferroni procedure tends to be rather conservative if the number of hypotheses is large or the test statistics are strongly positively correlated. Figure 2.3 displays the actual Type I error rate of the Bonferroni procedure in multiple testing problems with  $m = 2$  and 5 comparisons when the error rate is controlled at the one-sided 0.025 level. The test statistics are assumed to be equally correlated and follow a multivariate normal distribution. The common correlation coefficient is denoted by  $\rho$  ( $-1 < \rho \leq 1$  in the two-dimensional case and  $-1/4 < \rho \leq 1$  in the five-dimensional case). The probability of a Type I error is evaluated under the global null hypothesis (all hypotheses are true) based on 1,000,000 simulation runs. With  $m = 2$  comparisons, the error rate is very close to the nominal level when  $\rho \leq 0.3$  and becomes severely deflated when the test statistics are strongly positively correlated ( $\rho \geq 0.8$ ). In the case of  $m = 5$  comparisons, the actual error rate is below 0.02 even when the test statistics are moderately positively correlated ( $\rho$  is around 0.6).

## 2.6.2 Holm procedure and its extensions

The Holm (1979) procedure is a popular multiple testing procedure that demonstrates the advantages of a stepwise testing method.

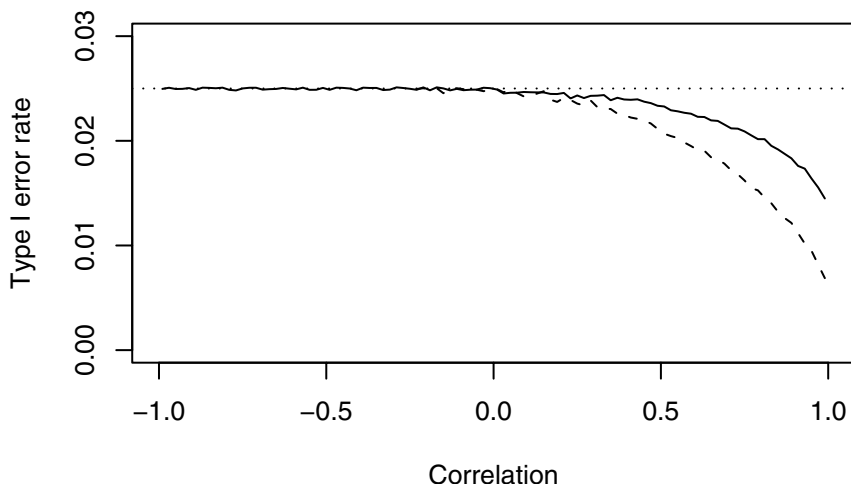
Assume first that the hypotheses are equally weighted. The Holm procedure is a step-down procedure that starts with the hypothesis associated with the most significant  $p$ -value and rejects it if the  $p$ -value is no greater than  $\alpha/m$ . If the first ordered hypothesis is rejected, the Holm procedure examines the next hypothesis in the sequence and so on. In general, this procedure is based on the following algorithm:

- Step 1. If  $p_{(1)} \leq \alpha/m$ , reject  $H_{(1)}$  and go to the next step. Otherwise retain all hypotheses and stop.
- Steps  $i = 2, \dots, m-1$ . If  $p_{(i)} \leq \alpha/(m-i+1)$ , reject  $H_{(i)}$  and go to the next step. Otherwise retain  $H_{(i)}, \dots, H_{(m)}$  and stop.
- Step  $m$ . If  $p_{(m)} \leq \alpha$ , reject  $H_{(m)}$ . Otherwise retain  $H_{(m)}$ .

This stepwise procedure is more powerful than the single-step Bonferroni procedure because it begins at the same significance level as the Bonferroni procedure ( $\alpha/m$ ) and tests the other hypotheses at successively higher levels.

---

<sup>1</sup>Although the Bonferroni inequality is named after the Italian mathematician Carlo Emilio Bonferroni, it is worth noting that Bonferroni's research focused on refining this inequality that actually goes back to the work of the British mathematician George Boole.



**FIGURE 2.3:** Type I error rate of the Bonferroni procedure under the global null hypothesis as a function of the number of comparisons and correlation (solid curve,  $m = 2$  comparisons, correlation  $> -1$ ; dashed curve,  $m = 5$  comparisons, correlation  $> -0.25$ ). The Bonferroni procedure is carried out at the one-sided 0.025 level. The dotted line is drawn at 0.025.

In the case of unequally weighted hypotheses, the Holm procedure can be defined based on the algorithms proposed by Holm (1979) or Benjamini and Hochberg (1997).

In order to prove that the Holm procedure controls the FWER, one can use a direct approach, as was done by Holm (1979), or utilize the closure principle introduced in Section 2.3.3. It is easy to show that, under the free combination condition, the Holm procedure is, in fact, a closed testing procedure in which each intersection hypothesis is tested using a global test based on the Bonferroni procedure. Since the Bonferroni procedure cannot reject more hypotheses than the Holm procedure, the latter serves as an example of how the power of a multiple testing procedure can be enhanced in a uniform manner by applying the closure principle. Note, however, that the Holm procedure rejects the global hypothesis if and only if the Bonferroni procedure does and therefore the conclusions regarding the conservative nature of the Bonferroni procedure (see Figure 2.3) also apply to the Holm procedure.

To compare the performance of the Holm procedure to that of the Bonferroni procedure, consider Scenario 1 in the dose-finding trial example (Sec-

tion 2.5.3). The ordered  $p$ -values are given by

$$p_{(1)} = p_4, \quad p_{(2)} = p_3, \quad p_{(3)} = p_2, \quad p_{(4)} = p_1.$$

At the first step of the Holm procedure, the smallest  $p$ -value,  $p_{(1)}$ , is compared with  $\alpha/4 = 0.00625$ . Since  $p_{(1)} < 0.00625$ , the corresponding hypothesis of no treatment effect is rejected and thus Dose D4 is superior to Placebo. Next,  $p_{(2)}$  is compared to  $\alpha/3 = 0.00833$  and the associated hypothesis is rejected (Dose D3 is superior to Placebo). Next,  $p_{(3)}$  is greater than the critical value,  $\alpha/2 = 0.0125$ , and therefore the Holm procedure retains the remaining two hypotheses (Doses D1 and D2 do not separate from Placebo).

## Shaffer procedure

Shaffer (1986) introduced a version of the Holm procedure for multiple testing problems with logical dependencies. Logical dependencies exist when truth of a subset of hypotheses necessarily implies truth of some other hypotheses. The most common example of logical dependencies occurs with all pairwise comparisons. For example, in the context of the dose-finding trial example introduced in Section 2.5.3, let  $\mu_0$  denote the mean improvement in the placebo group and  $\mu_i$  denote the mean improvement in the  $i$ th dose group,  $i = 1, 2, 3, 4$ . The hypotheses of interest are defined as follows:

$$H_{ij} : \mu_i = \mu_j.$$

Suppose that  $H_{12}$  and  $H_{13}$  are true, then it logically follows that  $H_{23}$  is also true. On the other hand, there are no logical dependencies when dose-placebo comparisons are considered. If  $H_{01}$  and  $H_{02}$  are true,  $H_{03}$  is not necessarily true.

When there are logical dependencies among the hypotheses, the divisors  $(m-i+1)$  in the Holm procedure may be replaced by divisors  $k_i$ , where  $k_i$  is the maximum number of the hypotheses  $H_{(i)}, \dots, H_{(m)}$  that can be simultaneously true given that  $H_{(1)}, \dots, H_{(i-1)}$  are false. Specifically, the Shaffer procedure rejects the hypothesis  $H_{(i)}$ ,  $i = 1, \dots, m$ , at the  $i$ th step if

$$p_{(j)} \leq \frac{\alpha}{k_j}, \quad j = 1, \dots, i.$$

In the dose-finding trial example introduced in Section 2.5.3, there are 10 pairwise comparisons of interest and thus  $k_1 = 10$ . This means that the critical value for the smallest  $p$ -value,  $p_{(1)}$ , is equal to that used by the Holm procedure, i.e.,  $\alpha/10$ . However, at the second step  $k_2 = 6$ . This represents a substantial improvement over the remaining number of comparisons, i.e.,  $(10 - 2 + 1) = 9$ , needed for the Holm procedure at the second step. Note that, when there are no logical dependencies among hypotheses, the Shaffer procedure reduces to the regular Holm procedure.

Shaffer developed two methods, Method 1 (described above) and Method

2, which uses the sequence of hypotheses  $H_{(1)}, \dots, H_{(m)}$  corresponding to the specific ordered  $p$ -values  $p_{(1)}, \dots, p_{(m)}$  observed in the study. The divisors  $l_i$  for Shaffer's Method 2 satisfy  $l_i \leq k_i$ ; hence Method 2 is uniformly more powerful than Method 1. Like the Holm procedure, the Method 2 procedure is a type of closed testing procedure based on the Bonferroni procedures for each intersection hypothesis. The procedure is set up as follows: closed testing is performed, in sequence, for  $H_{(1)}, \dots, H_{(m)}$  and testing stops at the first non-significant outcome. Thus Shaffer's Method 2 is called a *truncated closed testing procedure*. Truncation ensures that the procedure is monotone; i.e.,  $H_{(j)}$  cannot be rejected if  $H_{(i)}$  is not rejected and  $i < j$  (see Westfall and Tobias (2007) for details).

Shaffer's method has been modified recently to account for dependence structures as noted by Westfall and Tobias (2007); see [Section 2.7.3](#). This class of methods is especially useful for pairwise comparisons, which are uncommon in Phase III clinical trials but used more frequently in early-phase studies, and in general for comparisons that are logically intertwined, such as for non-pairwise comparisons using trend tests applied to dose-response analyses (see Chapter 3).

### 2.6.3 Fixed-sequence procedure

The fixed-sequence testing approach (Maurer et al., 1995; Westfall and Krishen, 2001) has found a variety of applications in clinical trials due to its straightforward stepwise form. The fixed-sequence procedure assumes that the order in which the hypotheses are tested,  $H_1, \dots, H_m$ , is pre-specified (this order normally reflects the clinical importance of the multiple analyses). Testing begins with 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. In other words, the hypothesis  $H_i$ ,  $i = 1, \dots, m$ , is rejected at the  $i$ th step if

$$p_j \leq \alpha, \quad j = 1, \dots, i.$$

The fixed-sequence procedure controls the FWER because, for each hypothesis, testing is conditional upon rejecting all hypotheses earlier in the sequence.

To demonstrate how a fixed-sequence strategy can be used in a clinical study, we will use the dose-finding trial example described in [Section 2.5.3](#). It may be reasonable to order the doses from D4 (highest dose) to D1 (lowest dose) since the higher doses are generally more likely to produce a significant treatment effect than the lower doses. In all three scenarios defined in [Section 2.5.3](#), the fixed-sequence procedure starts with the D4-Placebo comparison, proceeds to the next comparison in the sequence if the D4-Placebo statistic is significant at the one-sided 0.025 level and so on. Consider, for example, Scenario 1. Since all  $p$ -values are significant at 0.025 in this scenario, the fixed-sequence procedure rejects all hypotheses of no treatment effect.

### 2.6.4 Fallback procedure

The Holm and fixed-sequence procedures described in Sections 2.6.2 and 2.6.3 represent two different approaches to carrying out multiple testing procedures. In the case of the Holm procedure, testing is performed in a data-driven order. By contrast, the fixed-sequence procedure uses an *a priori* specified testing sequence. A compromise between the two testing approaches can be achieved by utilizing the fallback procedure introduced by Wiens (2003) and further studied by Wiens and Dmitrienko (2005), Dmitrienko, Wiens and Westfall (2006), Hommel, Bretz and Maurer (2007), Hommel and Bretz (2008) and Bretz et al. (2009b).

To introduce the fallback procedure, suppose the hypotheses  $H_1, \dots, H_m$  are ordered and allocate the overall error rate  $\alpha$  among the hypotheses according to their weights  $w_1, \dots, w_m$  (we will consider the general version of this procedure because it was designed specifically for the case of unequal weights). Specifically, the amount of the overall error rate assigned to  $H_i$  is equal to  $\alpha w_i$ ,  $i = 1, \dots, m$ . This process is similar to allocating the overall  $\alpha$  among the hypotheses in the weighted Bonferroni procedure. The fallback procedure is carried out as follows:

- Step 1. Test  $H_1$  at  $\alpha_1 = \alpha w_1$ . If  $p_1 \leq \alpha_1$ , reject this hypothesis; otherwise retain 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 retained. If  $p_i \leq \alpha_i$ , reject  $H_i$ ; otherwise retain 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 retained. If  $p_m \leq \alpha_m$ , reject  $H_m$ ; otherwise retain it.

It is instructive to compare the fallback and Holm procedures. Unlike the Holm procedure, the fallback procedure can continue testing even if a non-significant outcome is encountered by utilizing the fallback strategy (this explains why it is called the *fallback* procedure). If a hypothesis is retained, the next hypothesis in the sequence is tested at the level that would have been used by the weighted Bonferroni procedure. It was shown by Wiens and Dmitrienko (2005) that the fallback procedure is a closed testing procedure and thus it controls the FWER in the strong sense.

The fallback procedure is uniformly more powerful than the weighted Bonferroni procedure based on the same set of weights. In addition, the fallback procedure simplifies to the fixed-sequence procedure when  $w_1 = 1$  and  $w_2 = \dots = w_m = 0$ . Wiens and Dmitrienko (2005) suggested that the fallback procedure can be thought of as a compromise between the fixed-sequence and Hommel procedures (the Hommel procedure will be introduced in Section 2.6.8).

It was pointed out in Section 2.3.4 that the fallback procedure is not  $\alpha$ -exhaustive; i.e., when it is cast as a closed testing procedure, not all intersection hypotheses are tested at the full  $\alpha$  level. This means that one can construct a procedure that is uniformly more powerful than the fallback procedure and maintains the FWER at the same level. Wiens and Dmitrienko (2005) discussed several approaches to defining extended fallback procedures of this kind. In Section 2.6.5 we describe further properties of the regular and extended fallback procedures and discuss graphical tools for their visualization.

Using Scenario 1 in the dose-finding trial example (Section 2.5.3), we will demonstrate how to apply the fallback procedure to a multiple testing problem with equal weights (the four dose-placebo tests are equally weighted, i.e.,  $w_1 = \dots = w_4 = 1/4$ ). The fallback procedure begins with the comparison of Dose D4 to placebo and tests the associated hypothesis  $H_4$  at  $\alpha/4 = 0.0063$ . This hypothesis is rejected since  $p_4 < 0.0063$ . The  $\alpha$  level at which  $H_4$  was tested is carried over to the next hypothesis in the sequence,  $H_3$  (D3-Placebo comparison). This hypothesis is tested at  $2\alpha/4 = 0.0125$ . Note that  $p_3 < 0.0125$  and thus the fallback procedure rejects  $H_3$ , which means that  $\alpha/2$  is carried over to  $H_2$  (D2-Placebo comparison). This hypothesis is rejected at  $3\alpha/4 = 0.0188$  and the last hypothesis in the sequence,  $H_1$ , is tested at  $4\alpha/4 = 0.025$ . This hypothesis is also rejected by the fallback test.

## 2.6.5 Bonferroni-based closed testing procedures

In this section we show that the multiple testing procedures described in Sections 2.6.1–2.6.4 are all closed testing procedures based on the (weighted) Bonferroni test and thus follow the same construction principle. Understanding the closure principle (Section 2.3.3) enables one to take full advantage of its flexibility and to tailor the multiple testing procedure to the study objectives. In the following we will

- describe the class of Bonferroni-based closed testing procedures;
- give a sufficient characterization to derive sequentially rejective multiple testing procedures and demonstrate that many common procedures are in fact special cases thereof;
- construct simultaneous confidence intervals for procedures in the class;
- provide graphical tools that facilitate the derivation and communication of Bonferroni-based closed testing procedures based on sequentially rejective rules that are tailored to study objectives.

Because this section provides a general perspective of the methods described previously, the description is slightly more technical. In order to keep this section at a reasonable size, we omit the technical details and refer to the original publications instead.



## Class of Bonferroni-based closed testing procedures

As before, consider the problem of testing  $m$  hypotheses  $H_1, \dots, H_m$  and let  $I = \{1, \dots, m\}$  denote the associated index set. Recall from Section 2.3.3 that applying the closure principle leads to consideration of the intersection hypotheses  $H_J = \bigcap_{j \in J} H_j$ . For each intersection hypothesis  $H_J$  we assume a collection of non-negative weights  $w_j(J)$  such that they sum to 1, that is,  $0 \leq w_j(J) \leq 1$  and  $\sum_{j \in J} w_j(J) = 1$ . These weights quantify the relative importance of the hypotheses  $H_j$  included in the intersection  $H_J$ . As before, let  $p_j$  denote the raw  $p$ -value for  $H_j$ ,  $j \in I$ .

In this section we assume that each intersection hypothesis is tested with a weighted Bonferroni test. Consequently, we obtain the multiplicity adjusted  $p$ -values

$$p_J = \min\{q_j(J) : j \in J\}$$

for the weighted Bonferroni test for  $H_J$ , where

$$q_j(J) = \begin{cases} p_j/w_j(J) & \text{if } w_j(J) > 0, \\ 1 & \text{if } w_j(J) = 0. \end{cases}$$

This defines Class  $\mathcal{B}$  of all closed testing procedures that use weighted Bonferroni tests for each intersection hypothesis. Any collection of weights subject to the constraints given above can be used and thus one can choose the weights and tailor the closed testing procedure to the given study objectives.

To illustrate this, consider the simple two-hypothesis problem from Section 2.3.3. Consider the intersection hypothesis  $H_J$  with  $J = \{1, 2\}$  and associated weights  $w_1(J) = w_2(J) = 1/2$ . This results in the regular Bonferroni test and the adjusted  $p$ -value  $p_J = 2 \min(p_1, p_2)$ . If  $H_{\{1,2\}} = H_1 \cap H_2$  is rejected, so is either  $H_1$  or  $H_2$ , since they are tested subsequently at level  $\alpha$ . In other words, if  $H_{\{1,2\}}$  is rejected (the smaller of the two  $p$ -values is less than  $\alpha/2$ ), the remaining elementary hypothesis is tested at level  $\alpha$ , which is exactly the Holm procedure described in Section 2.6.2. Similarly, one can show that the Shaffer procedure (Section 2.6.2), fixed-sequence procedure (Section 2.6.3), fallback procedure (Section 2.6.4) and all Bonferroni-based gatekeeping procedures (Chapter 5) are examples of multiple testing procedures from Class  $\mathcal{B}$  (Hommel, Bretz and Maurer, 2007).

## Sequentially rejective Bonferroni-based closed testing procedures

It can further be shown that under a mild monotonicity condition on the weights  $w_j(J)$  the closure principle leads to powerful consonant multiple testing procedures (see Section 2.3.4 for the definition of consonance). Short-cut versions can thus be derived, which substantially simplify the implementation and interpretation of the related procedures. Hommel, Bretz and Maurer (2007) showed that all the procedures mentioned previously (with the notable

exception for the Shaffer procedure) belong to a subclass  $\mathcal{S} \subset \mathcal{B}$  of shortcut procedures characterized by the property

$$w_j(J) \leq w_j(J') \text{ for all } J' \subseteq J \subseteq I \text{ and } j \in J'.$$

This condition ensures that if an intersection hypothesis  $H_J$  is rejected, there is an index  $j \in J$ , such that  $p_j/w_j(J) \leq \alpha$  and the corresponding elementary hypothesis  $H_j$  can be rejected immediately by the closed testing procedure. Therefore, short-cut procedures of order  $m$  can be constructed; i.e., instead of testing  $2^m - 1$  hypotheses (as usually required by the closure principle), it is sufficient to test the elementary hypotheses  $H_1, \dots, H_m$  in  $m$  steps. This simplification is a key characterization of the Holm procedure and the results from Hommel, Bretz and Maurer (2007) ensure that this remains true for *any* procedure in  $\mathcal{S}$ . As a consequence, shortcut procedures from  $\mathcal{S}$  can be carried out with the following  $m$ -step procedure. Start testing the global intersection hypothesis  $H_I, I = \{1, \dots, m\}$ . If it is rejected, there is an index  $i \in I$  as described above such that  $H_i$  is rejected by the closed testing procedure. At the next step, one continues testing the global intersection  $H_{I \setminus i}$  of the remaining, not yet rejected hypotheses, and so on, until the first non-rejection.

## Simultaneous confidence intervals

The previous characterization for Class  $\mathcal{S}$  can also be used to construct compatible simultaneous confidence intervals introduced in Section 2.4.2 (Strassburger and Bretz, 2008; Guilhaud, 2008). Consider the one-sided null hypotheses  $H_i : \theta_i \leq \delta_i, i \in I = \{1, \dots, m\}$ , where  $\theta_1, \dots, \theta_m$  are the parameters of interest and  $\delta_1, \dots, \delta_m$  are pre-specified constants (e.g., noninferiority margins). Let  $\alpha_j(J) = \alpha w_j(J)$  denote the local significance levels with  $j \in J \subseteq I$ . Further, let  $L_i(\bar{\alpha})$  denote the marginal lower confidence limit for  $\theta_i$  at level  $1 - \bar{\alpha}, 1, \dots, m$ . Finally, let  $R$  denote the index set of hypotheses rejected by a multiple testing procedure from  $\mathcal{S}$ . Then, lower one-sided confidence limits for  $\theta_1, \dots, \theta_m$  with coverage probability of at least  $1 - \alpha$  are given by

$$\tilde{L}_i = \begin{cases} \delta_i & \text{if } i \in R \text{ and } R \neq I, \\ L_i(\bar{\alpha}_i) & \text{if } i \notin R, \\ \max(\delta_i, L_i(\bar{\alpha}_i)) & \text{if } R = I, \end{cases}$$

where  $\bar{\alpha}_i = \alpha_i(I \setminus R)$  if  $i \notin R \neq I$ . In the case  $R = I$ , where all hypotheses are rejected, the choice of the local levels  $\bar{\alpha}_i = \alpha_i(\emptyset)$  is arbitrary (Strassburger and Bretz, 2008). Thus, in order to compute the simultaneous confidence limits, one needs to know only the set  $R$  of rejected hypotheses and the corresponding local levels  $\bar{\alpha}_i$  for all indices  $i$  of retained hypotheses. Note that if not all hypotheses are rejected ( $R \neq I$ ), the confidence limits associated with the rejected hypotheses ( $i \in R$ ) essentially reflect the test decision  $\theta_i > \delta_i$  and the confidence limits associated with the retained hypotheses are the marginal

confidence limits at level  $\alpha_i(I \setminus R)$ . This method will be used to derive simultaneous confidence intervals for the Bonferroni, Holm, fixed-sequence and fallback procedures in Section 2.6.11.

## Graphical visualization

It was shown above that Class  $\mathcal{S}$  includes a variety of Bonferroni-based testing procedures, such as fixed-sequence, fallback and gatekeeping procedures. Using procedures in this class, one can map the difference in importance as well as the relationship between various study objectives onto a suitable multiple test procedure. However, since the procedures are based on the closure principle, one needs to specify the weights  $w_j(J)$  for each of the  $2^m - 1$  intersection hypotheses  $H_J, J \subseteq I$ . Unless these weights follow some simple and well-known specification rules (such as, for example, in the Holm procedure), the underlying test strategy may be difficult to communicate to clinical trial teams.

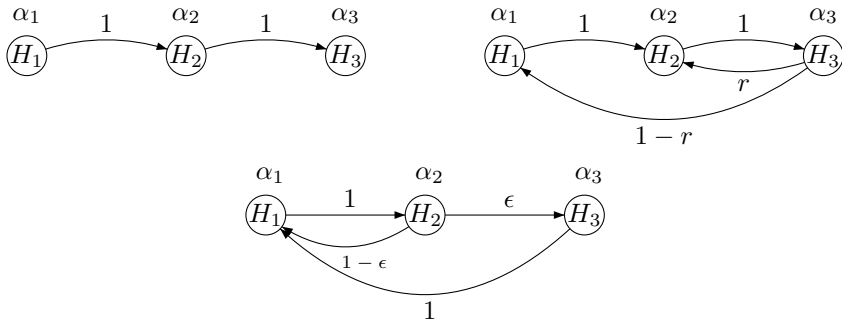
Graphical tools have been proposed instead, which help visualizing different sequentially rejective test strategies and thus to best tailor a multiple testing procedure to given study objectives (Bretz et al., 2009b). Using a graphical approach, the hypotheses  $H_1, \dots, H_m$  are represented by vertices with associated weights denoting the local significance levels  $\alpha_1, \dots, \alpha_m$ . The weight associated with a directed edge between any two vertices indicates the fraction of the (local) significance level that is shifted if the hypothesis at the tail of the edge is rejected.

To help illustrate this concept, consider a problem with three hypotheses  $H_1, H_2$  and  $H_3$ . The top left panel in Figure 2.4 displays a graphical representation of the fallback procedure introduced in Section 2.6.4. Each of the hypotheses is assigned an associated local significance level  $\alpha_i$ , such that  $\alpha_1 + \alpha_2 + \alpha_3 = \alpha$ . If  $H_1$  is rejected, then the level  $\alpha_1$  is carried over to  $H_2$ , as indicated by the edge pointing from  $H_1$  to  $H_2$ . If  $H_2$  is rejected at its local significance level (either  $\alpha_2$  or  $\alpha_1 + \alpha_2$ ), then that level is carried over to  $H_3$ , as indicated by the edge pointing from  $H_2$  to  $H_3$ .

It is important to note that graphical tools of this kind also help derive other, potentially more powerful testing strategies. Returning to the top left panel in Figure 2.4, one can see that, if  $H_3$  is rejected, its local significance level is not carried over to any other hypothesis. As shown in the top right panel, this significance level can be re-used by adding two further edges pointing back to  $H_1$  and  $H_2$ , where  $r = \alpha_2/(\alpha_1 + \alpha_2)$ . The resulting testing procedure is equivalent to the  $\alpha$ -exhaustive extension of the fallback procedure considered by Wiens and Dmitrienko (2005). Further, the bottom panel in Figure 2.4 displays yet another extension of the fallback procedure by shifting the significance level to the first hypothesis in the hierarchy that was not rejected so far (Hommel and Bretz, 2008). Here,  $\epsilon$  denotes an infinitesimally small weight indicating that the significance level is carried over from  $H_2$  to  $H_3$  only if both  $H_1$  and  $H_2$  are rejected. The motivation for this extension is that  $H_1$

is deemed more important than  $H_3$ . Thus, once  $H_2$  is rejected, its associated significance level should be carried over first to  $H_1$  before continuing to testing  $H_3$ .

We refer to Bretz et al. (2009b) for a detailed description of these ideas along with further extensions and examples, including a description of algorithms to derive the (updated) weights, simultaneous confidence intervals and adjusted  $p$ -values.



**FIGURE 2.4:** Graphical illustration of the fallback procedure (top left) and two extensions.

## 2.6.6 Reverse fixed-sequence procedure

The fixed-sequence procedure introduced in Section 2.6.3 has a sequentially rejective structure in the sense that each hypothesis can be rejected only if all previously examined hypotheses are rejected. The reverse fixed-sequence procedure is a mirror image of the regular fixed-sequence procedure.

Consider  $m$  ordered hypotheses  $H_1, \dots, H_m$ . The reverse fixed-sequence procedure begins with the first hypothesis in the sequence,  $H_1$ , and tests it at the full  $\alpha$  level. If the hypothesis is rejected, the other hypotheses are automatically rejected and the testing algorithm terminates. Otherwise,  $H_1$  is retained and the next hypothesis is tested. At the  $i$ th step of the algorithm, the procedure retains  $H_1, \dots, H_{i-1}$  and rejects  $H_i, \dots, H_m$  if  $p_j > \alpha$  for  $j = 1, \dots, i-1$  and  $p_i \leq \alpha$ ,  $i = 2, \dots, m$ . This procedure controls the FWER in the strong sense if  $H_1, \dots, H_m$  form a sequence of nested hypotheses; i.e.,  $H_i$  is a subset of  $H_j$  if  $i > j$ . Therefore, rejection of  $H_j$  implies rejection of all  $H_i$ 's for  $i > j$ .

To compute adjusted  $p$ -values for the reverse fixed-sequence procedure, note that  $p_i \geq p_j$  if  $i > j$  when the hypotheses are nested (in other words, it is easier to reject  $H_j$  compared to  $H_i$  when  $i > j$ ). Therefore, the adjusted  $p$ -value for  $H_i$  is simply equal to  $p_i$ ; i.e.,  $\tilde{p}_i = p_i$ ,  $i = 1, \dots, m$ .

Nested hypotheses are encountered in clinical trials with noninferiority and

superiority objectives. Specifically, consider a trial for testing the efficacy of a treatment versus a control with respect to a single endpoint. Let  $\delta$  denote an appropriate measure of the treatment difference (for example, the mean difference if the endpoint is continuous) and assume that a positive treatment difference indicates improvement. The trial's sponsor is interested in testing the hypotheses of noninferiority and superiority. The noninferiority hypothesis is defined as

$$H_1 : \delta \leq -\gamma,$$

where  $\gamma$  is a pre-specified positive noninferiority margin. The superiority hypothesis is given by:

$$H_2 : \delta \leq 0.$$

The two hypotheses can be tested sequentially using the fixed-sequence procedure:

- Begin with the noninferiority test and test  $H_1$  at the  $\alpha$  level.
- If noninferiority is established ( $H_1$  is rejected), switch to the superiority test, which is also carried out at the  $\alpha$  level.

This sequentially rejective procedure was described, among others, by Morikawa and Yoshida (1995). Morikawa and Yoshida pointed out that the reverse fixed-sequence procedure can also be applied in this problem due to the fact that  $H_1$  is a subset of  $H_2$ . The reverse fixed-sequence procedure is carried out as follows:

- Begin with the superiority test at the  $\alpha$  level.
- If superiority cannot be established ( $H_2$  is not rejected), carry out the noninferiority test at the  $\alpha$  level.

### 2.6.7 Simes global test

In this and the next two sections we will introduce the Simes global test and multiple testing procedures derived from the Simes test, including the Hochberg, Rom and Hommel procedures that will be defined later in this section.

The Simes test (Simes, 1986) focuses on testing the global hypothesis of no treatment effect; i.e.,

$$H_I = \bigcap_{i=1}^m H_i.$$

It rejects  $H_I$  if

$$p_{(i)} \leq i\alpha/m \text{ for at least one } i = 1, \dots, m,$$

where  $p_{(1)} < \dots < p_{(m)}$  are the ordered  $p$ -values.

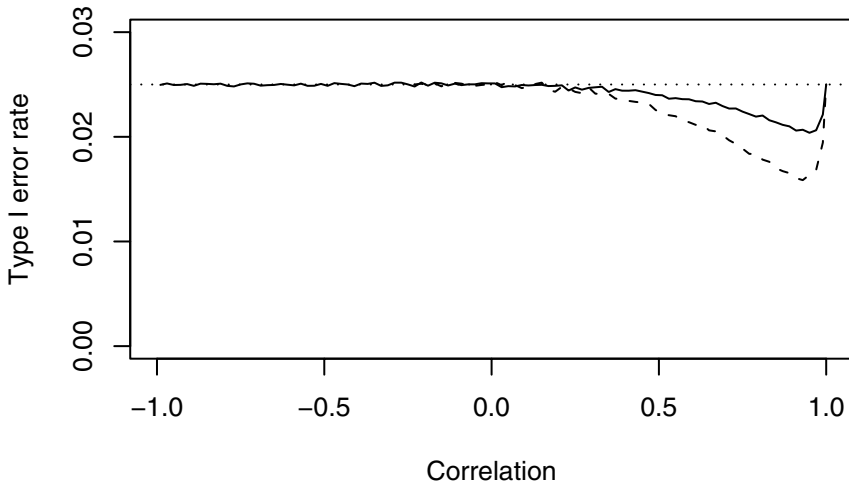
Note that the Simes test makes use of all ordered  $p$ -values (not just the smallest  $p$ -value) to test the global hypothesis and thus it is more powerful than a global test based on the Bonferroni procedure. It is also important to note that, unlike the Bonferroni procedure, the Simes test cannot be directly used to test the individual hypotheses  $H_1, \dots, H_m$ . In particular, one cannot reject  $H_{(i)}$  if  $p_{(i)} \leq i\alpha/m$ ,  $i = 1, \dots, m$ , since the FWER is not controlled in this case (Hommel, 1988).

Simes proved that this global test is exact in the sense that its size equals  $\alpha$  if  $p_1, \dots, p_m$  are independent. Since the assumption of independence is unlikely to be met in practice, several authors examined operating characteristics of this test under dependence. Hommel (1988) showed that the use of the Simes test can lead to an inflated Type I error probability. However, the worst-case scenario considered by Hommel corresponds to an extreme case that is very unlikely to be encountered in pharmaceutical applications. Hochberg and Rom (1995) examined the Type I error rate of the Simes test in the case of negatively-correlated normal variables and employed a simulation study to demonstrate that the Type I error rate is slightly inflated (about 10% inflation in the worst case). Samuel-Cahn (1996) showed via simulations that the Simes test preserves the Type I error rate in a one-sided setting with positively-correlated test statistics and in a two-sided setting regardless of the correlation. However, the test becomes anticonservative in the one-sided case if the test statistics are negatively correlated. Sarkar and Chang (1997) and Sarkar (1998) proved that the test protects the Type I error rate when the joint distribution of the test statistics exhibit a certain type of positive dependence; i.e., when the joint distribution is multivariate totally positive of order two (Karlin and Rinott, 1980)<sup>2</sup>. They showed that this condition is met in studies with multiple treatment-control comparisons under normal assumptions.

Figure 2.5 depicts the relationship between the Type I error rate of the Simes test, number of comparisons  $m = 2, 5$  and common correlation coefficient  $\rho$  in the case of normally distributed test statistics. The calculations are performed under the same assumptions as in Figure 2.3. Comparing the Type I error rate to that of the Bonferroni procedure (see Figure 2.3), it is easy to see that the Simes test performs better than the Bonferroni procedure under both weak and strong positive correlations. For example, the error rate in Figure 2.5 is only slightly below the nominal 0.025 level for both  $m = 2$  and  $m = 5$  comparisons when  $\rho \leq 0.4$  and is close to 0.025 when  $\rho$  approaches 1.

---

<sup>2</sup>It is worth pointing out that positive dependence is a more restrictive condition than positive correlation. However, in the case of a multivariate normal distribution the latter generally implies the former. For example, the positive dependence condition is satisfied when the test statistics follow a multivariate normal distribution with equal correlations and the common correlation is non-negative. In addition, this condition is satisfied in the multivariate normal case if all partial correlations are non-negative (Bolviken, 1982; Karlin and Rinott, 1983).



**FIGURE 2.5:** Type I error rate of the Simes test under the global null hypothesis as a function of the number of comparisons and correlation (solid curve,  $m = 2$  comparisons, correlation  $> -1$ ; dashed curve,  $m = 5$  comparisons, correlation  $> -0.25$ ). The Simes test is carried out at the one-sided 0.025 level. The dotted line is drawn at 0.025.

### 2.6.8 Hommel procedure

It was explained in Section 2.6.2 that the Holm procedure results from using a global test based on the Bonferroni for testing intersection hypotheses in a closed procedure. Similarly, the Hommel procedure (Hommel, 1988) results from using the Simes global test for testing individual intersection hypotheses. In the case of equally weighted hypotheses, the Hommel procedure can be applied using the following algorithm:

- Step 1. If  $p_{(m)} > \alpha$ , retain  $H_{(m)}$  and go to the next step. Otherwise reject all hypotheses and stop.
- Steps  $i = 2, \dots, m - 1$ . If  $p_{(m-j+1)} > (i - j + 1)\alpha/i$  for  $j = 1, \dots, i$ , retain  $H_{(m-i+1)}$  and go to the next step. Otherwise reject all remaining hypotheses and stop.
- Step  $m$ . If  $p_{(m-j+1)} > (i - j + 1)\alpha/i$  for  $j = 1, \dots, m$ , retain  $H_{(1)}$ ; otherwise reject it.

The Hommel procedure is easily extended to problems with unequally weighted hypotheses (Hommel, 1988). It protects the FWER under conditions that guarantee Type I error rate control for the Simes global test. It is uniformly more powerful than the Holm procedure because the Simes test is uniformly more powerful than the global test based on the Bonferroni procedure. For example, the Holm procedure rejects  $H_{(1)}$  if and only if  $p_{(1)} \leq \alpha/m$  whereas the Hommel procedure can reject this hypothesis when  $p_{(1)} > \alpha/m$ , e.g.,  $H_{(1)}$  is rejected if  $p_{(m)} \leq \alpha$ .

Using the dose-finding trial example from Section 2.5.3, we will illustrate the application of the stepwise algorithm defined above. Beginning with Scenario 1, note that all unadjusted  $p$ -values are less than 0.025, which immediately implies that the Hommel procedure rejects all hypotheses of no treatment effect. Now consider Scenario 2. The ordered  $p$ -values are given by

$$p_{(1)} = p_4, \quad p_{(2)} = p_3, \quad p_{(3)} = p_2, \quad p_{(4)} = p_1.$$

Since the largest  $p$ -value,  $p_{(4)}$ , is greater than 0.025, the Hommel procedure retains the hypothesis  $H_{(4)}$ . At the second step of the procedure,  $p_{(3)} > 0.025/2$  and  $p_{(4)} > 2(0.025/2)$ , which means that  $H_{(3)}$  is also retained. Further, the Hommel procedure retains  $H_{(2)}$  since

$$p_{(2)} > 0.025/3, \quad p_{(3)} > 2(0.025/3), \quad p_{(4)} > 3(0.025/3).$$

Finally, note that  $p_{(1)} \leq 0.025/4$  and thus  $H_{(1)}$  is rejected by the procedure.

### 2.6.9 Hochberg procedure

The Hochberg procedure (Hochberg, 1988) is another popular procedure based on the Simes global test. As was mentioned in Section 2.5.1, the Hochberg procedure is an example of a step-up procedure based on univariate  $p$ -values. Unlike step-down procedures (e.g., the Holm procedure), this procedure begins with the least significant  $p$ -value and examines the other  $p$ -values in a sequential manner until it reaches the most significant one.

Beginning with the case of equally weighted hypotheses, the decision rule for the Hochberg procedure is defined as follows:

- Step 1. If  $p_{(m)} > \alpha$ , retain  $H_{(m)}$  and go to the next step. Otherwise reject all hypotheses and stop.
- Steps  $i = 2, \dots, m-1$ . If  $p_{(m-i+1)} > \alpha/i$ , retain  $H_{(m-i+1)}$  and go to the next step. Otherwise reject all remaining hypotheses and stop.
- Step  $m$ . If  $p_{(1)} > \alpha/m$ , retain  $H_{(1)}$ ; otherwise reject it.

Extensions of the Hochberg procedure to the case of unequally weighted hypotheses were discussed in Tamhane and Liu (2008).

The Hochberg procedure controls the FWER under the same conditions



for which the Simes global test controls the Type I error rate. Further, this procedure is uniformly more powerful than the Holm procedure (Hochberg, 1988) but, on the other hand, it is uniformly less powerful than the Hommel procedure (Hommel, 1989).

Scenarios 1 and 2 in the dose-finding trial example (Section 2.5.3) will be used to illustrate the Hochberg procedure. Consider Scenario 1 and note that the Hochberg procedure is similar to the Hommel procedure in that it rejects all hypotheses if the largest unadjusted  $p$ -value is less than or equal to  $\alpha$ . Since the largest  $p$ -value is significant at the 0.025 level, all doses are significantly superior to Placebo after the Hochberg multiplicity adjustment. Further, it is easy to see that the Hochberg procedure finds only one significant dose-placebo comparison in Scenario 2. Note that

$$p_{(1)} = p_4, \quad p_{(2)} = p_3, \quad p_{(3)} = p_2, \quad p_{(4)} = p_1.$$

Therefore,

$$p_{(2)} > 0.025/3, \quad p_{(3)} > 0.025/2, \quad p_{(4)} > 0.025$$

and  $H_{(2)}$ ,  $H_{(3)}$  and  $H_{(4)}$  are retained. However,  $p_{(1)} \leq 0.025/4$  and thus the Hochberg procedure rejects  $H_{(1)}$ .

It is worth mentioning that the Hochberg procedure serves as a good example of the importance of a simple transparent structure in the choice of a multiple testing procedure. This procedure is very popular in clinical trial applications despite the fact that it is not  $\alpha$ -exhaustive and thus it can be improved in a uniform manner. In fact, the Hommel procedure is based on the full closure and consequently uniformly more powerful than the Hochberg procedure; however, the Hommel procedure is based on a more complicated algorithm. In addition, one can uniformly improve the power of the Hochberg procedure in the class of step-up procedures. For example, Rom (1990) derived a step-up procedure that is uniformly more powerful than the Hochberg procedure. The Rom procedure requires tabulation of critical values for ordered  $p$ -values whereas Hochberg's critical values do not require tabulation. Thanks to its computational simplicity, the Hochberg procedure remains the most popular Simes-based procedure used in practice.

### 2.6.10 Adjusted $p$ -values

This section discusses the computation of multiplicity-adjusted  $p$ -values for multiple testing procedures introduced in Sections 2.6.1–2.6.9 in the case of equally weighted hypotheses. Adjusted  $p$ -values are denoted by  $\tilde{p}_1, \dots, \tilde{p}_m$ . The hypothesis  $H_i$  is rejected if  $\tilde{p}_i \leq \alpha$ .

### Bonferroni procedure

The adjusted  $p$ -value for the hypothesis  $H_i$  is  $\tilde{p}_i = \min(1, mp_i)$ ,  $i = 1, \dots, m$ .

## Holm procedure

The adjusted  $p$ -values for the hypotheses  $H_{(1)}, \dots, H_{(m)}$  are defined sequentially in the following way:

$$\tilde{p}_{(i)} = \begin{cases} \min(1, mp_{(i)}) & \text{if } i = 1, \\ \max(\tilde{p}_{(i-1)}, (m - i + 1)p_{(i)}) & \text{if } i = 2, \dots, m. \end{cases}$$

## Fixed-sequence procedure

The adjusted  $p$ -value for  $H_i$  is given by  $\tilde{p}_i = \max(p_1, \dots, p_i)$ ,  $i = 1, \dots, m$ .

## Fallback and Hommel procedures

The adjusted  $p$ -values for the two procedures can be computed using the general method for closed testing procedures given in Section 2.4.1.

## Hochberg procedure

The adjusted  $p$ -values are defined recursively beginning with the largest  $p$ -value:

$$\tilde{p}_{(i)} = \begin{cases} p_{(i)} & \text{if } i = m, \\ \min(\tilde{p}_{(i+1)}, (m - i + 1)p_{(i+1)}) & \text{if } i = m - 1, \dots, 1. \end{cases}$$

## Dose-finding trial example

Table 2.2 displays adjusted  $p$ -values for the Bonferroni, Holm, fixed-sequence, fallback (assuming equally weighted hypotheses), Hommel and Hochberg procedures in the dose-finding trial example introduced in Section 2.5.3 (the unadjusted  $p$ -values are shown in Table 2.1).

The Bonferroni procedure rejects one hypothesis in Scenario 1 (Dose D4 is superior to Placebo) and also one hypothesis in Scenario 3 (Dose D3 is superior to placebo).

Since the Holm procedure is uniformly more powerful than the Bonferroni procedure, the Holm-adjusted  $p$ -values are no greater than the Bonferroni-adjusted  $p$ -values. This results in an additional significant test in Scenario 1 (D3-Placebo comparison) compared to the Bonferroni procedure. However, the numbers of hypotheses rejected by the Holm procedure in Scenarios 2 and 3 are the same as for the Bonferroni procedure.

The fixed-sequence procedure finds the following significant results: all doses are superior to Placebo in Scenario 1 and Doses D3 and D4 are superior to Placebo in Scenario 2. It is important to note that the fixed-sequence procedure outperforms the Bonferroni and Holm procedures in Scenarios 1 and 2

**TABLE 2.2:** Adjusted  $p$ -values for four dose-placebo tests in the dose-finding trial example under three scenarios. The asterisk identifies the adjusted  $p$ -values that are significant at the 0.025 level.

Procedure	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
Scenario 1				
Bonferroni	0.0912	0.0608	0.0284	0.0172*
Holm	0.0304	0.0304	0.0213*	0.0172*
Fixed-sequence	0.0228*	0.0152*	0.0071*	0.0043*
Fallback	0.0228*	0.0203*	0.0172*	0.0172*
Hommel	0.0228*	0.0228*	0.0213*	0.0142*
Hochberg	0.0228*	0.0228*	0.0213*	0.0172*
Scenario 2				
Bonferroni	0.1456	0.1188	0.0352	0.0280
Holm	0.0594	0.0594	0.0280	0.0280
Fixed-sequence	0.0364	0.0297	0.0088*	0.0070*
Fallback	0.0396	0.0396	0.0280	0.0280
Hommel	0.0364	0.0364	0.0264	0.0210*
Hochberg	0.0364	0.0364	0.0264	0.0264
Scenario 3				
Bonferroni	0.0648	0.0420	0.0220*	0.1316
Holm	0.0324	0.0315	0.0220*	0.0329
Fixed-sequence	0.0329	0.0329	0.0329	0.0329
Fallback	0.0220*	0.0220*	0.0220*	0.1316
Hommel	0.0324	0.0243*	0.0210*	0.0329
Hochberg	0.0324	0.0315	0.0220*	0.0329

when the true dose-response relationship is monotone. This illustrates an important property of the fixed-sequence procedure. The procedure maximizes the power when the (unknown) ordering of the hypotheses in terms of the true treatment effect is close to the selected ordering. However, if the monotonicity assumption is not met, the fixed-sequence procedure tends to perform poorly. For a discussion of the robustness of the fixed-sequence procedure with respect to the monotonicity assumption, see Section 4.3.

The fallback procedure rejects more hypotheses than the Holm procedure in Scenarios 1 and 3. Further, the fallback procedure is more robust to departures from the monotonicity assumption than the fixed-sequence procedure and leads to more rejections when the first dose in the pre-determined sequence does not separate from placebo. For instance, the fallback procedure detects three significant dose-placebo comparisons in Scenario 3 whereas the fixed-sequence procedure does not find any significant results.

The Hommel procedure rejects more hypotheses than the Holm procedure in Scenarios 1, 2 and 3 (note that the Hommel procedure always rejects all hypotheses if the largest unadjusted  $p$ -value is  $\leq \alpha$ ). However, it performs only as well as the Holm procedure in Scenario 3 when the dose-response curve

has an umbrella shape and finds fewer significant results than the fallback procedure.

The Hochberg procedure finds all dose-placebo comparisons significant in Scenario 1 and thus rejects more hypotheses of no treatment effect than the Holm procedure. However, unlike the Hommel procedure, the Hochberg procedure fails to detect the drug effect at Dose D4 in Scenario 2. Scenario 3 shows that the Hochberg procedure can sometimes be less powerful compared to the fallback procedure.

### 2.6.11 Simultaneous confidence intervals

In this section we will define simultaneous confidence intervals for Bonferroni-based procedures defined in Section 2.6.5; i.e., for the Bonferroni, Holm, fixed-sequence and fallback procedures, in problems with equally weighted hypotheses. Simultaneous confidence intervals for the Hommel and Hochberg procedures have not been explicitly defined in the multiple comparison literature.

To define simultaneous confidence intervals, consider a one-sided parametric multiple testing problem defined as follows. The hypothesis of no treatment effect

$$H_i : \theta_i \leq 0$$

is tested versus a one-sided alternative

$$K_i : \theta_i > 0,$$

where  $i = 1, \dots, m$  and  $\theta_1, \dots, \theta_m$  are parameters of interest, for example, mean treatment differences or differences in proportions. Let  $\hat{\theta}_i$  denote an estimate of  $\theta_i$  and assume that  $\hat{\theta}_i$  is normally distributed with mean  $\theta_i$  and standard deviation  $\sigma_i$ . The estimated standard error of  $\theta_i$  is denoted by  $s_i$ ,  $i = 1, \dots, m$ . Further,  $z_x$  denotes the  $(1 - x)$ -quantile of the standard normal distribution.

#### Bonferroni procedure

A one-sided  $100(1 - \alpha)\%$  confidence interval for  $\theta_1, \dots, \theta_m$  is given by  $(\tilde{L}_i, \infty)$ ,  $i = 1, \dots, m$ , where

$$\tilde{L}_i = \hat{\theta}_i - z_{\alpha/m} s_i.$$

#### Holm procedure

One-sided simultaneous confidence intervals for the Holm procedure were developed by Strassburger and Bretz (2008) and Guilbaud (2008). Based on the general results presented in Section 2.6.5, the lower limits of one-sided

100(1 -  $\alpha$ )% confidence intervals are given by

$$\tilde{L}_i = \begin{cases} 0 & \text{if } i \in R \text{ and } R \neq I, \\ \hat{\theta}_i - z_{\alpha/(m-r)} s_i & \text{if } i \notin R, \\ \max(0, \hat{\theta}_i - z_{\alpha/m} s_i) & \text{if } R = I, \end{cases}$$

where  $R$  is the index set of rejected hypotheses and  $r$  is the number of rejected hypotheses. Here the first case applies to the hypotheses rejected by the Holm procedure ( $i \in R$ ) when the procedure retains some of the hypotheses ( $R \neq I$ ). The second case applies to the hypotheses retained by the Holm procedure ( $i \notin R$ ) and the third case corresponds to scenarios when all hypotheses are rejected by the procedure ( $R = I$ ). It is worth noting that the lower limit for a parameter is set to 0 whenever the Holm procedure rejects the corresponding hypothesis of no treatment effect. The only exception is when the procedure rejects all hypotheses. In this case the lower limits associated with rejected hypotheses can be greater than 0.

### Fixed-sequence procedure

Hsu and Berger (1999) constructed simultaneous confidence intervals associated with the fixed-sequence procedure in one-sided parametric problems. Based on the general results from Section 2.6.5, we obtain the lower limits of the one-sided 100(1 -  $\alpha$ )% simultaneous confidence intervals for  $\theta_i$ ,  $i = 1, \dots, m$ ,

$$\tilde{L}_i = \begin{cases} 0 & \text{if } i \in R \text{ and } R \neq I, \\ \hat{\theta}_i - z_{\alpha} s_i & \text{if } i = i^* \text{ and } R \neq I, \\ \min_{i \in I} \{\hat{\theta}_i - z_{\alpha} s_i\} & \text{if } R = I, \end{cases}$$

where  $i^*$  denotes the first hypothesis in the sequence not rejected by the procedure when the procedure retains some of the hypotheses ( $R \neq I$ ) and  $R$  is the index set of rejected hypotheses. Note that the hypotheses  $H_j$  with  $j > i^*$  are not tested and therefore no confidence intervals are available for the associated parameters. Further, it follows from this definition that the lower limits of simultaneous confidence intervals for the fixed-sequence procedure are similar to the Holm-adjusted limits in the sense that they are set to 0 if the corresponding hypothesis is rejected unless the fixed-sequence procedure rejects all hypotheses.

### Fallback procedure

An extension of the method for setting up simultaneous confidence intervals for Bonferroni-based closed testing procedures proposed by Strassburger and Bretz (2008) can be used to define simultaneous confidence intervals for

the fallback procedure. The lower limits of one-sided  $100(1 - \alpha)\%$  confidence intervals are derived as follows. First, for any non-empty index set  $J \subseteq I$ , let

$$\alpha_i(J) = \begin{cases} 0 & \text{if } i \notin J, \\ \alpha(i - \ell_i(J))/m & \text{if } i \in J, \end{cases}$$

where  $\ell_i(J)$  is the largest index in  $J$  that is smaller than  $i$  if  $i$  is not the smallest index in  $J$  and  $\ell_i(J) = 0$  if  $i$  is the smallest index in  $J$ . Similarly, for any non-empty index set  $J \subseteq I$  and  $i \notin J$ , let

$$\alpha_i^*(J) = \frac{1}{m - |J|} \left( \alpha - \sum_{j \in J} \alpha_j(J) \right),$$

where  $|J|$  is the number of elements in  $J$ .

The lower limits are given by

$$\tilde{L}_i = \begin{cases} \min_{J \subseteq A} \max(0, \hat{\theta}_i - z_{\alpha_i^*(J)} s_i) & \text{if } i \in R \text{ and } R \neq I, \\ \hat{\theta}_i - z_{\alpha_i(A)} s_i & \text{if } i \in A, \\ \max(0, \hat{\theta}_i - z_{\alpha/m} s_i) & \text{if } R = I, \end{cases}$$

where  $A$  and  $R$  are the index sets of retained and rejected hypotheses, respectively. These lower limits take advantage of the fact that the fallback procedure is not  $\alpha$ -exhaustive and are uniformly sharper than those based on the general method presented in Section 2.6.5. Unlike the lower limits for the Holm and fixed-sequence procedures, the fallback-adjusted lower limits are not automatically set to 0 for parameters corresponding to rejected hypotheses of no treatment effect.

## Dose-finding trial example

Table 2.3 shows the lower limits of one-sided 97.5% simultaneous confidence intervals for the four mean treatment differences under Scenario 1 in the dose-finding trial example (the unadjusted lower limits are presented in Table 2.1). The limits are computed for the Bonferroni, Holm, fixed-sequence and fallback procedures (as before, the weights are assumed to be equal in the fallback procedure).

Table 2.3 illustrates key properties of simultaneous confidence intervals. First of all, comparing the lower limits displayed in Table 2.3 to the adjusted  $p$ -values presented in Table 2.2 under Scenario 1, it is easy to see that the lower limit is less than 0 if the procedure fails to reject the corresponding hypothesis of no treatment effect at the 0.025 level.

The Holm-adjusted lower limits are sharper than those for the Bonferroni procedure when the latter fails to reject a hypothesis of no treatment effect

**TABLE 2.3:** Lower limits of one-sided 97.5% simultaneous confidence intervals for the mean dose-placebo treatment differences in the dose-finding trial example (Scenario 1)

Procedure	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
Bonferroni	−0.71	−0.47	−0.05	0.20
Holm	−0.34	−0.10	0.00	0.00
Fixed-sequence	0.07	0.07	0.07	0.07
Fallback	0.00	0.00	0.00	0.20

(see the lower limits for the D1-Placebo, D2-Placebo and D3-Placebo tests). However, when both procedures reject a hypothesis, the Holm-adjusted lower limit is less informative than the Bonferroni-adjusted lower limit. Consider, for example, the D4-Placebo test. In this case the Bonferroni-adjusted lower limit is positive and thus provides information about the likely magnitude of the treatment difference whereas the Holm-adjusted lower limit is simply equal to 0.

Further, the lower limits for the fixed-sequence procedure are positive and constant across the four dose-placebo comparisons because the procedure rejects all hypotheses in Scenario 1. The fallback procedure also rejects all hypotheses and thus the associated lower limits are nonnegative. However, the first three lower limits are set to 0 and only one lower limit is positive (D4-Placebo test).

## 2.7 Parametric multiple testing procedures

In certain situations, for example, in dose-finding clinical trials with normally distributed outcomes, it is possible to improve the power of  $p$ -value-based procedures by taking advantage of parametric assumptions about the joint distribution of the test statistics. Multiple testing procedures that rely on these assumptions are known as parametric procedures. The most well-known parametric procedure is the Dunnett procedure (Dunnett, 1955) developed for problems with multiple dose-control comparisons. This single-step procedure is described in this section along with other parametric procedures such as the stepwise Dunnett procedures and parametric Shaffer procedure.

The following setting will be used throughout this section. Consider a dose-finding clinical trial designed to compare  $m$  doses or regimens of a treatment to a placebo. For simplicity, a balanced one-way layout will be assumed; i.e.,

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

where  $y_{ij}$  is the response of the  $j$ th patient in the  $i$ th treatment group,  $i =$

$0, \dots, m$  ( $i = 0$  denotes the placebo group) and  $j = 1, \dots, n$ . The errors,  $\varepsilon_{ij}$ ,  $i = 0, \dots, m$ ,  $j = 1, \dots, n$ , are normally distributed with mean 0 and common standard deviation  $\sigma$ .

The testing problem is formulated in terms of the  $m$  treatment-placebo comparisons; i.e., the hypotheses  $H_i : \theta_i = 0$ ,  $i = 1, \dots, m$ , are tested against the one-sided alternatives  $K_i : \theta_i > 0$ ,  $i = 1, \dots, m$ , where  $\theta_i = \mu_i - \mu_0$ . Let  $t_i$  be the  $t$  statistic for testing  $H_i$ ; i.e.,

$$t_i = \frac{\bar{y}_i - \bar{y}_0}{s\sqrt{2/n}},$$

where  $s$  is the pooled sample standard deviation.

### 2.7.1 Single-step Dunnett procedure

The single-step Dunnett procedure can be thought of as a set of two-sample tests for  $H_1, \dots, H_m$  adjusted for multiplicity. However, unlike  $p$ -value-based procedures described in Section 2.6, the Dunnett procedure is based on the joint distribution of the test statistics and thus accounts for the correlation among the test statistics.

It is easy to show that, under  $H_i$ ,  $t_i$  follows the standard univariate  $t$  distribution with  $\nu = 2(n - 1)$  df and thus the regular (unadjusted) critical value for  $t_i$  is  $t_\alpha(\nu)$ ; i.e., the  $(1 - \alpha)$ -quantile of the  $t$  distribution. Similarly, the Dunnett-adjusted critical value for  $t_1, \dots, t_m$ , denoted by  $u_\alpha(m, \nu)$ , is the  $(1 - \alpha)$ -quantile of the distribution of the maximum of  $t$ -distributed random variables with  $\nu = (m + 1)(n - 1)$  df. In other words,  $u_\alpha(m, \nu) = F^{-1}(1 - \alpha|m, \nu)$ , where  $F(x|m, \nu)$  is the cumulative distribution function of the one-sided Dunnett distribution; i.e.,

$$F(x|m, \nu) = P\{\max(t_1, \dots, t_m) \leq x\},$$

where the probability is evaluated under the overall null hypothesis  $H_0 : \theta_1 = \dots = \theta_m = 0$ . The Dunnett procedure rejects  $H_i$  if  $t_i \geq u_\alpha(m, \nu)$ ,  $i = 1, \dots, m$ .

Dunnett-adjusted critical values are smaller than Bonferroni-adjusted critical values. Therefore, the use of the Dunnett procedure leads to more powerful inferences compared to the Bonferroni procedure. As an illustration, we will use Scenario 1 in the dose-finding trial example introduced in Section 2.5.3. The Bonferroni-adjusted critical value in this problem is  $t_{\alpha/m}(\nu)$  with  $\alpha = 0.025$ ,  $m = 4$  and  $\nu = 2(n - 1) = 152$ , i.e., 2.53. As was stated in Section 2.6.1, only one test statistic is greater than this critical value (Dose D4 is superior to placebo). The Dunnett-adjusted critical value is given by  $u_\alpha(m, \nu)$  with  $\alpha = 0.025$ ,  $m = 4$  and  $\nu = (m + 1)(n - 1) = 380$ . The critical value is 2.45 and the Dunnett procedure detects two significant dose-placebo comparisons in Scenario 1 (Doses D3 and D4 are superior to placebo).

It is important to note that the Dunnett procedure can be applied to any problem in which the test statistics for multiple dose-placebo comparisons



asymptotically follow a multivariate normal distribution. For example, this procedure can be used in clinical trials with categorical outcomes provided the proportions are not so different as to cause serious heteroscedasticity problems (Chuang-Stein and Tong, 1995). However, it is generally preferable to fit the model desired and use the multiplicity adjustments that follow from that model specifically. Hothorn, Bretz and Westfall (2008) gave examples that include binary and other parametric non-normally distributed cases.

### 2.7.2 Stepwise Dunnett procedures

The Dunnett procedure defined in Section 2.7.1 is similar to the Bonferroni procedure in that it also has a single-step structure and is not  $\alpha$ -exhaustive (for this reason, the Dunnett procedure can be thought of as a parametric version of the Bonferroni procedure). This implies that one can develop more powerful parametric procedures by applying the closure principle. This section introduces two stepwise versions of the Dunnett procedure: a step-down procedure analogous to the Holm procedure and a step-up procedure analogous to the Hochberg procedure. Both procedures are uniformly more powerful than the single-step Dunnett procedure.

#### Step-down Dunnett procedure

A step-down procedure which serves as a parametric extension of the Holm procedure presented in Section 2.6.2 was developed by Naik (1975) and Marcus, Peritz and Gabriel (1976). Recall that the Holm procedure is defined using ordered  $p$ -values and, to define the step-down Dunnett procedure, we will use ordered test statistics  $t_{(1)} > \dots > t_{(m)}$  and associated hypotheses  $H_{(1)}, \dots, H_{(m)}$ .

The step-down Dunnett procedure is a sequentially rejective procedure that first assesses if there is sufficient evidence to reject  $H_{(1)}$  under the overall null hypothesis, i.e., all doses are no different from placebo. If  $H_{(1)}$  cannot be rejected, testing stops. Otherwise, the next hypothesis in the sequence,  $H_{(2)}$ , is tested under the assumption that the remaining  $m - 1$  hypotheses are true and so on.

Defining  $u_\alpha(i, \nu)$ ,  $i = 1, \dots, m$ , as the  $(1 - \alpha)$ -quantile of the  $i$ -variate  $t$  distribution with  $\nu = (m + 1)(n - 1)$  df, the step-down version of the Dunnett procedure is implemented using the following algorithm:

- Step 1. If  $t_{(1)} \geq c_1$ , where  $c_1 = u_\alpha(m, \nu)$ , reject  $H_{(1)}$  and go to the next step. Otherwise retain all hypotheses and stop.
- Steps  $i = 2, \dots, m - 1$ . If  $t_{(i)} \geq c_i$ , where  $c_i = u_\alpha(m - i + 1, \nu)$ , reject  $H_{(i)}$  and go to the next step. Otherwise retain  $H_{(i)}, \dots, H_{(m)}$  and stop.
- Step  $m$ . If  $t_{(m)} \geq c_m$ , where  $c_m = u_\alpha(1, \nu)$ , reject  $H_{(m)}$ . Otherwise retain  $H_{(m)}$ .

The step-down Dunnett procedure uses the critical value associated with the single-step Dunnett procedure at the first step, i.e.,  $c_1 = u_\alpha(m, \nu)$ . Further,  $c_1 > c_2 > \dots > c_m$  and thus the other hypotheses are tested using successively sharper critical values. This implies that the step-down procedure rejects as many (and potentially more) hypotheses than the single-step Dunnett procedure. In addition, the step-down Dunnett procedure is uniformly more powerful than the Holm procedure.

The step-down Dunnett procedure defined above assumes a balanced one-way layout. The step-down procedure in the general unbalanced case was considered by Bofinger (1987) and Dunnett and Tamhane (1991).

Scenario 1 in the dose-finding trial example given in Section 2.5.3 will be used to illustrate the step-down testing algorithm. The ordered  $t$  statistics in this scenario are given by

$$t_{(1)} = t_4, \quad t_{(2)} = t_3, \quad t_{(3)} = t_2, \quad t_{(4)} = t_1$$

and the critical values at Steps 1 through 4 are equal to 2.45, 2.36, 2.22 and 1.97, respectively. The first ordered test statistic,  $t_{(1)} = 2.64$ , is greater than the corresponding critical value, 2.45, and thus the hypothesis  $H_{(1)}$  is rejected (Dose D4 is superior to placebo). The next ordered statistic,  $t_{(2)} = 2.46$ , is compared to 2.36 and is again significant (Dose D3 is superior to placebo). However, the other two hypotheses of no treatment effect are retained since  $t_{(3)} = 2.17$  is less than 2.22.

## Step-up Dunnett procedure

A step-up version of the Dunnett procedure was proposed by Dunnett and Tamhane (1992). It is conceptually similar to the step-up Hochberg and Rom procedures described in Section 2.6.9.

The step-up testing algorithm is set up as follows. The ordered  $t$  statistics  $t_{(1)} > \dots > t_{(m)}$  are compared to suitably defined critical values  $c_1, \dots, c_m$  in a stepwise fashion starting with the least significant test statistic, i.e.,  $t_{(m)}$ . At each step, all remaining hypotheses are rejected if the test statistic is greater or equal to the corresponding critical value. Specifically, testing is performed as follows:

- Step 1. If  $t_{(m)} < c_1$ , retain  $H_{(m)}$  and go to the next step. Otherwise reject all hypotheses and stop.
- Steps  $i = 2, \dots, m - 1$ . If  $t_{(m-i+1)} < c_i$ , retain  $H_{(m-i+1)}$  and go to the next step. Otherwise reject all remaining hypotheses and stop.
- Step  $m$ . If  $t_{(1)} < c_m$ , retain  $H_{(1)}$  and reject it otherwise.

Dunnett and Tamhane (1992) showed that the step-up Dunnett procedure is uniformly more powerful than the single-step Dunnett procedure as well as

the Hochberg procedure which serves as an example of a nonparametric step-up procedure. However, the step-up Dunnett procedure does not uniformly dominate the step-down Dunnett procedure in terms of power. The step-up procedure tends to be more powerful than the step-down Dunnett procedure when most of the true mean treatment-control differences in a dose-finding study are positive.

The critical values in the step-up procedure are defined in such a way that the FWER is controlled at the  $\alpha$  level. The following recursive algorithm can be used to obtain the critical values in a balanced one-way layout. Let  $T_1, \dots, T_m$  be random variables with the same joint distribution as  $t_1, \dots, t_m$  under the global null hypothesis. The critical value  $c_1$  is found from

$$P(T_1 \geq c_1) = \alpha.$$

Further, given  $c_1, \dots, c_{i-1}$ , the critical value  $c_i$  is chosen so that

$$P(T_{(1)} \geq c_1 \text{ or } T_{(2)} \geq c_2 \text{ or } \dots \text{ or } T_{(i)} \geq c_i) = \alpha,$$

where  $T_{(1)} < \dots < T_{(i)}$ . Note that  $c_1$  is simply the  $(1 - \alpha)$ -quantile of the univariate  $t$  distribution with  $\nu = (m + 1)(n - 1)$  df and thus the step-up Dunnett procedure is similar to the Hochberg procedure in that it also rejects all hypotheses if the least significant  $p$ -value is no greater than  $\alpha$ .

Calculation of critical values for the step-up procedure in the general unbalanced case was considered by Dunnett and Tamhane (1995) and Grechanovsky and Pinsker (1999). An efficient algorithm for computing the critical values was proposed by Kwong and Liu (2000).

The step-up algorithm will be illustrated using Scenarios 1 and 2 in the dose-finding trial example from Section 2.5.3. To carry out the step-up Dunnett procedure, we will use the critical values given in Table 2 of Dunnett and Tamhane (1992) with  $\nu = \infty$ . The critical values are given by  $c_1 = 1.96$ ,  $c_2 = 2.22$ ,  $c_3 = 2.36$  and  $c_4 = 2.45$ .

Considering Scenario 1, the ordered test statistics are given by

$$t_{(1)} = t_4, \quad t_{(2)} = t_3, \quad t_{(3)} = t_2, \quad t_{(4)} = t_1.$$

At the first step of the algorithm, the least significant test statistic,  $t_{(4)} = 2.01$ , is compared to  $c_1$ . Since the test statistic is greater than the critical value, the step-up Dunnett procedure rejects all hypotheses of no treatment effect and thus all doses are declared superior to Placebo.

The ordered test statistics in Scenario 2 are again given by

$$t_{(1)} = t_4, \quad t_{(2)} = t_3, \quad t_{(3)} = t_2, \quad t_{(4)} = t_1.$$

Since the least significant test statistic,  $t_{(4)} = 1.80$ , is no greater than  $c_1$ , the step-up Dunnett procedure retains the hypothesis  $H_{(4)}$ . At the next step, the test statistic,  $t_{(3)} = 1.89$ , is less than the corresponding critical value  $c_2$  and thus  $H_{(3)}$  is also retained. Lastly,  $t_{(2)} = 2.38$  exceeds the critical value  $c_3$  and, as a consequence, the step-up Dunnett procedure rejects the two remaining hypotheses (Doses D3 and D4 are superior to Placebo).

### 2.7.3 Extended Shaffer-Royen procedure

Westfall and Tobias (2007) discussed the extended Shaffer-Royen procedure that serves as a parametric extension of Shaffer's Method 2 described in Section 2.6.2 to account for logical dependencies among hypotheses. When the hypotheses are formulated in terms of dose-placebo contrasts that are not logically related, the procedure reduces precisely to the step-down Dunnett method described above. More generally, the extended Shaffer-Royen procedure is a truncated closed testing procedure similar to the Shaffer's Method 2 procedure. Note that the latter procedure uses the Bonferroni test for each intersection hypothesis  $H_I$  while the parametric procedure uses the distribution of the maximum test statistic for the intersection hypothesis  $H_I = \cap_{i \in I} H_i$ , i.e.,  $\max_{i \in I} t_i$ , to make the method more powerful. "Royen" appears in the name of the procedure since Royen (1989) first applied it to the problem of testing all pairwise comparisons. The extended Shaffer-Royen procedure can be used to test arbitrary contrasts when the contrasts are logically related, as occurs, for example, when the multiple contrasts represent subgroups (see the example in Hochberg and Westfall, 2000).

### 2.7.4 Adjusted $p$ -values and simultaneous confidence intervals

In this section we will introduce algorithms for computing adjusted  $p$ -values and associated simultaneous confidence intervals for the Dunnett-based parametric procedures.

#### Single-step Dunnett procedure

The adjusted  $p$ -values for individual hypotheses are found using the multivariate  $t$  distribution. Specifically, the adjusted  $p$ -value for  $H_i$  is  $\tilde{p}_i = 1 - F(t_i|m, \nu)$ , where  $F(x|m, \nu)$  is the cumulative distribution function of the one-sided Dunnett distribution with  $\nu = (m+1)(n-1)$  defined in Section 2.7.1. In other words, the adjusted  $p$ -value for  $H_i$  is found from

$$t_i = u_{\tilde{p}_i}(m, \nu).$$

The lower limits of one-sided  $100(1-\alpha)\%$  simultaneous confidence intervals for the mean treatment differences  $\theta_i = \mu_i - \mu_0$ ,  $i = 1, \dots, m$ , are defined as follows:

$$\tilde{L}_i = \hat{\theta}_i - u_{\alpha}(m, \nu)s_i,$$

where  $s_i$  is the standard error of  $\hat{\theta}_i$ , i.e.,  $s_i = s\sqrt{2/n}$ .

#### Step-down Dunnett procedure

The adjusted  $p$ -values for the step-down Dunnett procedure are found using the following algorithm (Dunnett and Tamhane, 1992). First, define

$\gamma_1, \dots, \gamma_m$  as follows:

$$t_{(i)} = u_{\gamma_i}(m - i + 1, \nu), \quad i = 1, \dots, m,$$

where  $\nu = (m + 1)(n - 1)$ . The adjusted  $p$ -values are given by

$$\tilde{p}_{(i)} = \begin{cases} \gamma_i & \text{if } i = 1, \\ \max(\tilde{p}_{i-1}, \gamma_i) & \text{if } i = 2, \dots, m. \end{cases}$$

Simultaneous confidence intervals for the step-down procedure were derived by Bofinger (1987) and Stefansson, Kim and Hsu (1988). The lower limits of one-sided  $100(1 - \alpha)\%$  simultaneous confidence intervals for  $\theta_i = \mu_i - \mu_0$ ,  $i = 1, \dots, m$ , are derived using the Stefansson-Kim-Hsu method defined below. The underlying algorithm is similar to the algorithm used in the calculation of lower simultaneous confidence limits for the step-down version of the Bonferroni procedure, e.g., Holm procedure (see [Section 2.6.11](#)):

$$\tilde{L}_i = \begin{cases} 0 & \text{if } i \in R \text{ and } R \neq I, \\ \hat{\theta}_i - c_{r+1}s_i & \text{if } i \notin R, \\ \max(0, \hat{\theta}_i - c_m s_i) & \text{if } R = I, \end{cases}$$

where  $R$  is the index set of rejected hypotheses,  $I = \{1, \dots, m\}$  and  $r$  is the number of rejected hypotheses. As in [Section 2.6.11](#), the first case defines the lower limits for the hypotheses rejected by the step-down procedure ( $i \in R$ ) when some other hypotheses are retained ( $R \neq I$ ). Note that the lower limits for the treatment differences asserted to be significant by the step-down procedure ( $i \in R$ ) are automatically set to 0 (unless all hypotheses are rejected, i.e.,  $R = I$ ). When confidence limits for the step-down procedure are equal to 0, they may be less informative than positive limits for the treatment differences that are found significant by the single-step Dunnett procedure. Further, the second case applies to the hypotheses retained by the step-down procedure ( $i \notin R$ ) and the third case defines the lower limits when all hypotheses are rejected ( $R = I$ ).

## Step-up Dunnett procedure

Adjusted  $p$ -values for the step-up Dunnett procedure are defined in Dunnett and Tamhane (1992) and Grechanovsky and Pinsker (1999). The algorithm is computationally intensive and calculation of adjusted  $p$ -values for this procedure will not be discussed in this book. Further, a method for constructing simultaneous confidence intervals for the step-up Dunnett procedure has not been developed yet.

**TABLE 2.4:** Adjusted  $p$ -values for four dose-placebo tests in the dose-finding trial example under three scenarios. The asterisk identifies the adjusted  $p$ -values that are significant at the 0.025 level.

Procedure	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
Scenario 1				
Single-step Dunnett	0.0715	0.0493	0.0242*	0.0152*
Step-down Dunnett	0.0280	0.0280	0.0190*	0.0152*
Scenario 2				
Single-step Dunnett	0.1090	0.0909	0.0297	0.0238*
Step-down Dunnett	0.0535	0.0535	0.0238*	0.0238*
Scenario 3				
Single-step Dunnett	0.0523	0.0351	0.0191*	0.0994
Step-down Dunnett	0.0298	0.0278	0.0191*	0.0329

**TABLE 2.5:** Lower limits of one-sided 97.5% simultaneous confidence intervals for the mean dose-placebo treatment differences in the dose-finding trial example (Scenario 1).

Procedure	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
Single-step Dunnett	−0.64	−0.40	0.02	0.27
Step-down Dunnett	−0.31	−0.07	0	0

Dose-finding trial example

Table 2.4 lists adjusted  $p$ -values produced by the single-step and step-down Dunnett procedures in the dose-finding trial example (see [Section 2.5.3](#)). It is easy to verify that the adjusted  $p$ -values for the single-step Dunnett procedure are uniformly smaller than those associated with the Bonferroni procedure (Bonferroni-adjusted  $p$ -values are displayed in [Table 2.2](#)). The single-step Dunnett procedure finds two significant dose-placebo comparisons in Scenario 1, one significant comparison in Scenario 2 and one significant comparison in Scenario 3. The step-down Dunnett procedure is uniformly superior to the single-step Dunnett procedure as well as the nonparametric step-down procedure, i.e., Holm procedure.

Further, the lower limits of one-sided 97.5% simultaneous confidence intervals for the mean treatment differences in the dose-finding trial example are displayed in Table 2.5 (assuming Scenario 1). Note that the lower confidence limits for the single-step Dunnett procedure are uniformly sharper than those for the Bonferroni procedure ([Table 2.3](#)). Further, the lower confidence limits associated with the step-down Dunnett procedure are greater than the lower confidence limits for the single-step Dunnett procedure when the latter does not reject a hypothesis and, when both procedures reject a hypothesis, the opposite is true.

### 2.7.5 Multiple comparisons in general linear models

In this section we extend the single-step Dunnett procedure from Section 2.7.1 and describe a general approach to the problem of constructing multiple testing procedures in general linear models that account for the stochastic dependencies among the test statistics. The general theory is covered, among others, by Hochberg and Tamhane (1987), Hsu (1996, Chapter 7) and Bretz, Hothorn and Westfall (2008).

Consider the common linear model

$$Y = X\beta + \varepsilon,$$

where  $Y$  is an  $n \times 1$  response vector,  $X$  is a fixed and known  $n \times p$  design matrix,  $\beta$  is an unknown  $p \times 1$  parameter vector and  $\varepsilon$  is an  $n \times 1$  vector of independent normally distributed errors with mean 0 and unknown variance  $\sigma^2$ . The least square unbiased estimates of  $\beta$  and  $\sigma$  are given by

$$\hat{\beta} = (X'X)^-X'Y \text{ and } s^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{\nu},$$

respectively, where  $\nu = n - \text{rank}(X)$  is the error degrees of freedom and  $(X'X)^-$  is some generalized inverse of  $X'X$ .

Let  $C$  denote a constant  $p \times m$  matrix which describes the experimental questions of interest. Each column  $c_i$ ,  $i = 1, \dots, m$ , of  $C$  defines a single experimental comparison of interest. Without loss of generality, we consider the associated one-sided testing problem

$$H_i : \theta_i \leq 0, \quad i = 1, \dots, m,$$

where  $\theta_i = c_i'\beta$ .

The hypotheses are tested using the test statistics

$$t_i = \hat{\theta}_i/s_i, \quad i = 1, \dots, m,$$

where  $s_i = s\sqrt{c_i'(X'X)^-c_i}$ ,  $i = 1, \dots, m$ .

It can be shown that the joint distribution of  $t_1, \dots, t_m$  is multivariate  $t$  with  $\nu$  degrees of freedom and correlation matrix  $DC'(X'X)^-CD$ , where  $D = \text{diag}(c_i'(X'X)^-c_i)^{-1/2}$ . In the asymptotic case  $\nu \rightarrow \infty$  or if  $\sigma$  is known, the corresponding multivariate normal distribution can be used instead. Let  $u_\alpha$  denote the critical value derived from the multivariate normal or  $t$  distribution. Then,  $H_i$  is rejected if  $t_i \geq u_\alpha$ . Equivalently, adjusted  $p$ -values  $\tilde{p}_i$  can be calculated from the multivariate normal or  $t$  distribution and we reject  $H_i$  if  $\tilde{p}_i \leq \alpha$ . Numerical integration methods to calculate the multivariate normal and  $t$  probabilities required for the computation of critical values and adjusted  $p$ -values are described by Genz and Bretz (2002, 2009). Finally, one-sided simultaneous confidence intervals for  $\theta_1, \dots, \theta_m$  with simultaneous coverage probability  $1 - \alpha$  are given by

$$(-\infty, \hat{\theta}_i + u_\alpha s_i], \quad i = 1, \dots, m,$$

To illustrate this framework, we revisit the single-step Dunnett test considered in Section 2.7.1 for comparing  $m$  treatments with a control. Here,  $p = m + 1$  and  $\beta = (\mu_0, \dots, \mu_m)'$  is the parameter vector. The index  $i = 0$  denotes the placebo control to which the remaining  $m$  treatment arms are compared. The associated  $C$  matrix is

$$C_{m+1 \times m} = \begin{pmatrix} -1 & -1 & \dots & -1 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Thus, for the  $i$ th vector

$$c'_i = (-1, 0, \dots, 1, 0, \dots, 0),$$

we obtain the pairwise mean treatment differences  $\theta_i = \mu_1 - \mu_0$ ,  $i = 1, \dots, m$ . It is easy to show that the resulting  $m$  treatment-control comparisons are exactly those considered in Section 2.7.1.

This general framework includes many standard parametric multiple testing procedures beyond the Dunnett procedure, such as the Tukey procedure for all pairwise comparisons, the trend tests of Williams (1971) and Marcus (1976) (see Chapter 3) and other procedures listed in Bretz, Genz and Hothorn (2001). The framework enables one to include covariates and/or factorial treatment structures in classical regression and ANOVA applications. An extension to more general parametric and semi-parametric models relying on standard asymptotic multivariate normal assumptions was provided by Hothorn, Bretz and Westfall (2008), which allows a unified treatise of multiple comparisons for generalized linear models, mixed models, survival models, etc. Note that in this section we focused on single-step procedures. Using the results from Section 2.7.2, more powerful closed testing procedures can be constructed based on the parametric framework described here. These procedures account for the stochastic dependencies among the test statistics and exploit inherent logical constraints for general parametric models. These methods are all implemented in the *multcomp* package reviewed in Section 2.9 and described in Bretz, Hothorn and Westfall (2010).

---

## 2.8 Resampling-based multiple testing procedures

Resampling is a general term that encompasses the bootstrap, permutation analysis, and parametric simulation-based analyses. The general method extends parametric methods described in Section 2.7.

The resampling-based methods



- make fewer assumptions about the data-generating process (for example, normality does not have to be assumed) and yield more robust multiple testing procedures,
- utilize data-based distributional characteristics, including discreteness and correlation structure which yield more powerful procedures.

At the same time, one needs to be aware of some drawbacks of the resampling-based approach. In particular, resampling-based methods

- are often approximate, requiring large sample sizes and/or simulations to justify their use (except for permutation-based methods, which are exact even for small sample sizes; see [Section 2.8.4](#)),
- can be computationally difficult,
- are based on careful modeling of the data-generating process requiring very complex models in some cases such as survival analysis.

This section begins with the description of general principles used in the construction of resampling-based multiple testing procedures (Section 2.8.1). Sections 2.8.2–2.8.3 give examples of parametric and non-parametric resampling-based procedures and Section 2.8.4 discusses exact permutation-based procedures.

### 2.8.1 Closed resampling-based procedures

The closure principle introduced in Section 2.3.3 provides a convenient, flexible and powerful foundation to describe resampling-based multiple testing procedures. Consider  $m$  hypotheses of interest denoted by  $H_1, \dots, H_m$ . As in Section 2.7, let  $T_1, \dots, T_m$  denote the random variables representing the test statistics associated with these hypotheses and  $t_1, \dots, t_m$  denote the observed test statistics (realizations of  $T_1, \dots, T_m$ ).

In general, any  $\alpha$ -level test may be used to test intersection hypotheses in the closed family induced by  $H_1, \dots, H_m$ . Previous sections have discussed the use of procedures based on the minimum  $p$ -value (e.g., Bonferroni procedure) or maximum test statistic (e.g., Dunnett procedure). However, there are many other choices. In an ANOVA setting arising in dose-finding studies,  $F$ -statistics may be used to test intersection hypotheses. Similarly, in multi-center, subgroup, or other types of analyses where it is desired to combine data across locations, Fisher combination tests or other meta-analytic tests may be used. The choice of test statistic to use should primarily be based on power considerations. Once a powerful test statistic is chosen, resampling can be used to ensure that the test is robust to violations of distributional and/or dependence assumptions.

While power is the main concern for choice of a test statistic, expediency becomes important when the number of hypotheses  $m$  is large. There are

$2^m - 1$  intersection hypotheses in the closed family and, if  $m$  is large, it is computationally impossible to test every single intersection. However, the computational burden can be eased dramatically if the following assumptions are made:

- For each non-empty index set  $I \subseteq \{1, \dots, m\}$ , the intersection hypothesis  $H_I = \cap_{i \in I} H_i$  is tested using the maximum statistic  $t_{\max}(I) = \max_{i \in I} t_i$ .
- The *subset pivotality condition* (Westfall and Young, 1993) is met, i.e., for each non-empty index set  $I$ , the distribution of  $T_{\max}(I)$  under  $H_I$  is identical to the distribution of  $T_{\max}(I)$  under the global null hypothesis  $H_N$ .

The benefit of the two assumptions is that stepwise resampling-based procedures similar to the Holm and stepwise Dunnett procedures can be constructed. In other words, we need to test only  $m$  hypotheses corresponding to the ordered test statistics  $t_{(1)} > \dots > t_{(m)}$  rather than all  $2^m - 1$  intersection hypotheses. Further, resampling can be done simultaneously under the global null hypothesis rather than separately for each intersection hypothesis.

To illustrate, let  $H_{(1)}, \dots, H_{(m)}$  denote the hypotheses associated with the ordered test statistics. The step-down resampling-based procedure is defined as follows:

- Step 1. Reject  $H_{(1)}$  if

$$P(\max(T_1, \dots, T_m) \geq t_{(1)}) \leq \alpha$$

and go to the next step. Otherwise retain  $H_{(1)}, \dots, H_{(m)}$  and stop.

- Steps  $i = 2, \dots, m - 1$ . Reject  $H_{(i)}$  if

$$P(\max(T_i, \dots, T_m) \geq t_{(i)}) \leq \alpha$$

and go to the next step. Otherwise retain  $H_{(i)}, \dots, H_{(m)}$  and stop.

- Step  $m$ . Reject  $H_{(m)}$  if

$$P(T_m \geq t_{(m)}) \leq \alpha.$$

Otherwise retain  $H_{(m)}$  and stop.

The probabilities in this algorithm are computed under the global null hypothesis. The step-down procedure is identical to the full-blown closed testing procedure based on the examination of each individual intersection hypothesis in the closed family. Adjusted  $p$ -values are the probabilities shown at each step, adjusted for monotonicity as in the case of the parametric step-down Dunnett procedure.

It is worth noting that, in general, we do not need to use resampling to implement this step-down procedure. This will be the case when the probabilities at Steps 1 through  $m$  can be computed directly. However, when direct computations are not feasible, resampling methods are used to obtain these probabilities.

In addition to the use of the maximum test statistics and subset pivotality condition, we also need to assume that there are no logical constraints among the hypotheses to be able to construct a step-down procedure identical to the original closed testing procedure. If there are logical constraints, power can be improved by restricting attention only to intersection hypotheses consistent with the constraints. However, in this case the computational shortcuts disappear and we are back in the position of having to evaluate the tests for all intersection hypotheses. The step-down procedure defined above can still be used, though, as it provides a conservative approximation to the closed testing procedure.

Sections 2.8.2–2.8.4 illustrate the general step-down method in some special cases.

## 2.8.2 Step-down Dunnett procedures based on parametric and nonparametric resampling

The step-down Dunnett procedure for parametric problems arising in dose-finding trials was presented in Section 2.7.2. We will show in this section that the step-down procedure defined in Section 2.8.1 is obtained (at least in the simulation limit) via parametric normal resampling. In addition, we will show how to extend this method simply using bootstrap resampling for cases where the normality assumption is violated.

Consider the setting introduced in Section 2.7. In particular, assume that the responses follow the ANOVA model defined in that section and assume that the errors are independent, identically distributed random variables with mean 0 and variance  $\sigma^2$ . Further, consider the same set of hypotheses, i.e.,  $H_i : \mu_i = \mu_0, i = 1, \dots, m$ , and define the closed family associated with  $H_1, \dots, H_m$ . The intersection hypothesis  $H_I$  is tested using the test statistic  $t_{\max}(I) = \max_{i \in I} t_i$ , where  $t_i$  is the  $t$ -statistic for comparing the  $i$ th group to placebo. The  $p$ -value for this intersection hypothesis is given by

$$p_I = P(T_{\max}(I) \geq t_{\max}(I)),$$

where  $t_{\max}(I)$  is defined based on the observed test statistics  $t_1, \dots, t_m$  and the probability is computed under the global null hypothesis.

If the errors in the ANOVA model are normally distributed, we can use the multivariate  $t$  distribution to calculate the  $p$ -value for each intersection hypothesis. In this case, the step-down procedure introduced in Section 2.8.1 will simplify to the step-down Dunnett procedure. Alternatively, as shown below, the same step-down algorithm results exactly (in the simulation limit) from parametric resampling and these parametric resampling procedures suggest

natural extensions to nonparametric resampling (also known as bootstrap-based procedures).

## Parametric resampling

To set up a step-down procedure based on parametric resampling, consider the ANOVA model with normally distributed errors and note that the distribution of  $T_i$ ,  $i \in I$ , does not depend on  $\mu_j$  for  $j \notin I$  when  $H_I$  is true, i.e., subset pivotality holds. Therefore, we can simulate  $T_i$  by parametrically resampling the data as follows:

- Step 1. Generate a resampled data set

$$y_{ij}^* = 0 + \varepsilon_{ij}^*, \quad i = 0, \dots, m, \quad j = 1, \dots, n,$$

where the  $\varepsilon_{ij}^*$  are independent, identically distributed normal variables with mean 0 and variance  $s^2$ . Actually, one can use any variance, since the distribution of  $T_i$  is also free of  $\sigma^2$ . However, use of  $s^2$  clarifies the connection to the nonparametric resampling algorithm described later in this section.

- Step 2. Compute the 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 computed from the resampled data set.

Repeat the two steps  $B$  times ( $B$  needs to be large, e.g.,  $B = 100,000$ ). The probability for the  $i$ th step in the step-down algorithm is approximately (within binomial simulation error) the proportion of the  $B$  samples where

$$\max(T_i^*, \dots, T_m^*) \geq t_{(i)}, \quad i = 1, \dots, m.$$

## Nonparametric resampling

The parametric resampling method generalizes very easily to models in which the distribution of random errors is unknown. However, one does not simply “resample the data and hope for the best.” Instead care is needed to specify a model and resampling scheme that is appropriate for the given model.

Consider again the ANOVA model and assume that the random errors follow a general distribution with mean zero and finite variance, rather than a normal distribution. This model is called a *location-shift model*. As before, define the closed family associated with  $H_1, \dots, H_m$ . The  $p$ -value for the intersection hypothesis  $H_I$  is again given by

$$p_I = P(T_{\max}(I) \geq t_{\max}(I)),$$

**TABLE 2.6:** Adjusted  $p$ -values produced by the bootstrap-based step-down Dunnett procedure for four dose-placebo tests in the dose-finding trial example under three scenarios. The asterisk identifies the adjusted  $p$ -values that are significant at the 0.025 level.

Scenario	Test			
	D1-Placebo	D2-Placebo	D3-Placebo	D4-Placebo
1	0.0279	0.0279	0.0189*	0.0151*
2	0.0534	0.0534	0.0238*	0.0238*
3	0.0299	0.0278	0.0191*	0.0330

where the probability is computed under the global null hypothesis based on the true distribution  $F$ . What is different is that the  $p$ -value now depends on the underlying distribution  $F$ , which is unknown. If we knew  $F$ , we could use the parametric resampling algorithm described above with the  $\varepsilon_{ij}^*$ 's simulated from this distribution. Since  $F$  is unknown, its estimate, denoted by  $\hat{F}$ , will be used and the  $\varepsilon_{ij}^*$ 's will be generated from  $\hat{F}$  in the algorithm. The resulting  $p$ -value is only approximate, since  $\hat{F}$  is not equal to the true distribution  $F$ . However, typically (though not automatically) with larger sample sizes,  $\hat{F}$  becomes closer to  $F$ . Still, in any finite-sample multiple testing problem, the fact that the procedure is approximate means that simulation studies are generally needed to assess the adequacy of the approximation.

There are a variety of ways to obtain  $p$ -values in the nonparametric algorithm, including various flavors of simple and smoothed bootstraps. As an example, consider the procedure based on the basic bootstrap method. It is virtually identical to the parametric resampling-based procedure defined above. The only difference is that the  $\varepsilon_{ij}^*$ 's are sampled with replacement from the sample random errors  $e_{ij} = y_{ij} - \bar{y}_i$ ,  $i = 0, \dots, m$ ,  $j = 1, \dots, n$ . Note that, as in the parametric case, the distribution of  $T_i$ ,  $i \in I$ , is free of  $\mu_j$ ,  $j \notin I$ , under  $H_I$ , so one can simulate data with mean 0, i.e., subset pivotality holds again. The probabilities in the nonparametric algorithm are approximated by the proportions of the  $B$  samples in which

$$\max(T_i^*, \dots, T_m^*) \geq t_{(i)}, \quad i = 1, \dots, m.$$

To illustrate the bootstrap-based step-down procedure (i.e., the procedure based on nonparametric resampling), consider the dose-finding trial example given in Section 2.5.3. Adjusted  $p$ -values produced by the bootstrap-based procedure with  $B = 5,000,000$  bootstrap samples are shown in Table 2.6. These adjusted  $p$ -values differ little from the adjusted  $p$ -values produced by the parametric step-down Dunnett procedure (see [Table 2.4](#)).

## Bootstrap-based simultaneous confidence intervals

To develop the simultaneous confidence intervals for a general nonparametric case, consider again the ANOVA model in which the errors  $\varepsilon_{ij}$  are a random sample from an unspecified distribution  $F$  having finite variance  $\sigma^2$ . Further, consider the lower limits of one-sided simultaneous confidence intervals for the mean treatment differences  $\theta_i = \mu_i - \mu_0$ ,  $i = 1, \dots, m$ , associated with the single-step Dunnett procedure defined in Section 2.7.4. The lower limit for  $\theta_i$  is given by  $\tilde{L}_i = \hat{\theta}_i - u_\alpha s_i$ , where  $u_\alpha$  is the Dunnett-adjusted critical value. The analog to this critical value is  $u_\alpha(F)$ , where

$$P(\theta_i \geq \hat{\theta}_i - u_\alpha(F) s_i \text{ for all } i) = 1 - \alpha,$$

or, equivalently,

$$P(\max(T_1, \dots, T_m) \leq u_\alpha(F)) = 1 - \alpha.$$

Note that the joint distribution of  $T_1, \dots, T_m$  does not depend on the parameters  $\mu_1, \dots, \mu_m$ .

Since  $F$  is unknown, the critical value  $u_\alpha(F)$  is estimated using a bootstrap method, e.g., using the basic bootstrap method defined above. Note that, since  $F$  is estimated, the resulting critical values are doubly approximate  $\hat{u}_\alpha(\hat{F})$ , with approximation due to simulation error resulting from  $B$  simulations (this error can be reduced with greater  $B$ ) and due to the approximation of  $F$  via  $\hat{F}$  (this error is reduced by increasing the sample size  $n$ ).

As an illustration, the one-sided 0.025-level bootstrap-based critical value for Scenario 1 in the dose-finding trial example from Section 2.5.3 is 2.45 (this value was computed using  $B = 1,000,000$  bootstrap samples). The bootstrap-based critical value is equal to the Dunnett critical value  $u_{0.025}(4, 380) = 2.45$ , shown in Section 2.7.1, and the bootstrap-based simultaneous confidence intervals are identical to those displayed in Table 2.5.

While the example shows no difference between the parametric and resampling-based approaches, there are cases where resampling matters. These include multiple endpoints, where resampling provides a convenient way to incorporate correlation structure, and adverse events, where sparseness is automatically incorporated to allow much greater power.

### 2.8.3 Step-down resampling-based procedures for multivariate linear models

Nonparametric resampling-based procedures (bootstrap-based procedures) are more compelling in complex models where there are no standard methods for handling non-normal data and/or complex dependence structure. As in Westfall and Young (1993), consider the general multivariate regression model

$$Y = X\beta + \varepsilon,$$

where  $Y$  is a  $n \times v$  matrix of response variables,  $X$  is a full rank  $n \times b$  design matrix,  $\beta$  is a  $b \times v$  matrix of regression parameters, and  $\varepsilon$  is a  $n \times v$  matrix of random error terms, all with mean zero. This model subsumes basic ANOVA as well as analysis of covariance models arising in clinical trials when univariate ANOVA models are expanded to include covariates, e.g., demographic and clinical characteristics. Note that  $n$  does not have to be greater than  $v$ , so the model can be used also for gene expression data where  $v$  is typically much larger than  $n$ .

Assume that the rows  $\varepsilon_i$  of  $\varepsilon$  are independent and identically distributed according to some multivariate distribution  $F$ . Specific tests of interest in such models are usually one-dimensional (or single-degree-of-freedom) tests which may be formulated with respect to hypotheses defined as

$$H_i : \theta_i = 0, \quad i = 1, \dots, m,$$

where  $\theta_i = c_i' \beta d_i$ ,  $i = 1, \dots, m$ , and  $c_i$  and  $d_i$  are vectors of constants that specify the hypotheses of interest.

Commonly-used test statistics and unadjusted  $p$ -values for each  $H_i$  are obtained from the matrix of least squares estimates

$$\hat{\beta} = (X'X)^{-1}X'Y$$

and the usual unbiased error variance-covariance matrix estimate

$$S = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n - b).$$

The statistic  $T_i = \hat{\theta}_i/s_i$ , where

$$\hat{\theta}_i = c_i' \hat{\beta} d_i \text{ and } s_i = \sqrt{d_i' S d_i c_i' (X'X)^{-1} c_i},$$

is commonly used to test  $H_i$ . When the error distribution  $F$  is multivariate normal and  $H_i$  is true,  $T_i$  has the  $t$  distribution with  $n - b$  degrees of freedom.

Note that when  $H_i$  is true,

$$T_i = \frac{c_i' \hat{\beta} d_i}{\sqrt{d_i' S d_i c_i' (X'X)^{-1} c_i}} = \frac{c_i' (X'X)^{-1} X' \varepsilon d_i}{\sqrt{d_i' S d_i c_i' (X'X)^{-1} c_i}},$$

showing that subset pivotality holds. Note also that

$$S = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n - b} = \frac{\varepsilon' \{I - X(X'X)^{-1}X'\} \varepsilon}{n - b},$$

so the distribution of the test statistics is completely determined by the distribution of  $\varepsilon$ , the design matrix and the chosen comparisons.

It is desirable to account for the correlations between the variables in the multiple comparisons procedure. This can be done via parametric resampling, where error vectors are simulated from a multivariate normal distribution with estimated covariance matrix, or by nonparametric bootstrap-based resampling of error vectors. A simple bootstrap algorithm to test the intersection hypothesis  $H_I$  is as follows:

- Step 1. Generate a resampled data set

$$Y^* = 0 + \varepsilon^*,$$

where the rows  $\varepsilon_i^*$  of  $\varepsilon^*$  are chosen with replacement from the sample error vectors  $\{e_i^*\}$ , where  $e_i^*$  is the  $i$ th row of the sample error matrix  $e = Y - X\hat{\beta}$ . Because the test statistics do not depend on  $\beta$  under  $H_I$ , there is no need to include  $X\hat{\beta}$  in the resampled data set, unlike other bootstrap procedures.

- Step 2. Compute the statistics  $\hat{\beta}^* = (X'X)^{-1}X'Y^*$ ,  $S^* = (Y^* - X\hat{\beta}^*)'(Y^* - X\hat{\beta}^*)/(n - b)$  and

$$T_i^* = \frac{c_i' \hat{\beta}^* d_i}{\sqrt{d_i' S^* d_i c_i' (X'X)^{-1} c_i}}.$$

Steps 1 and 2 are repeated  $B$  times. The bootstrap  $p$ -value  $\hat{p}_I$  is defined as the proportion of the  $B$  samples in which  $T_{\max}^*(I) \geq t_{\max}(I)$ , where  $T_{\max}^*(I) = \max_{i \in I} T_i^*$  and  $t_{\max}(I) = \max_{i \in I} t_i$ . Again, as described in Section 2.8.1, testing can be performed sequentially, based on ordered observed test statistics.

As an example, consider the Phase II clinical trial described in Westfall et al. (1999, Section 11.3). In this trial, the efficacy of a treatment was evaluated using four endpoints labeled Y1 through Y4 (Endpoint Y4 is reverse-scored). Table 2.7 displays the raw and adjusted  $p$ -values produced by the bootstrap-based procedure with  $B = 1,000,000$  bootstrap samples and Holm procedure (again, the adjusted  $p$ -values are monotonicity-enforced, as with the step-down Dunnett procedure). The main competitor for the bootstrap-based step-down procedure is the Holm step-down procedure, but the latter loses power because it does not account for correlations. Unlike the previous examples with the Dunnett procedure where there was virtually no difference between the bootstrap method and the parametric counterpart, a clear advantage of the bootstrap method is seen in this example.

## 2.8.4 Permutation-based procedures

Permutation-based resampling provides an ideal method for constructing multiple testing procedures with multivariate two-sample data, including multiple endpoints and adverse events, since

- the conditions giving rise to subset pivotality can be relaxed relative to the location-shift models presented above,
- the resulting multiple comparisons method is exact, with no finite-sample error incurred from estimating  $F$  via  $\hat{F}$ ,



**TABLE 2.7:** One-sided raw  $p$ -values and adjusted  $p$ -values produced by the Holm and bootstrap-based procedures for Endpoints Y1–Y4. The asterisk identifies the adjusted  $p$ -values that are significant at the 0.025 level.

Endpoint	Raw $p$ -value	Adjusted $p$ -value	
		Holm	Bootstrap
Y1	0.0060	0.0242*	0.0186*
Y2	0.0071	0.0242*	0.0186*
Y3	0.0993	0.0993	0.0988
Y4	0.0095	0.0242*	0.0186*

- the resulting methods can be exceptionally more powerful than parametric counterparts when data are sparse, in particular when the data are binary.

Consider the following two-sample problem. Let  $Y_i = (Y_{i1}, \dots, Y_{in})$  denote the multivariate  $v$ -dimensional data vectors in the  $i$ th sample,  $i = 1, 2$ . One might assume the general location-shift model

$$Y = X\beta + \varepsilon$$

described in Section 2.8.3, where the matrix  $X$  has two columns of indicator (dummy) variables, and where the  $v$ -dimensional rows of  $\varepsilon$  are independent identically distributed variables, but this is somewhat restrictive. Instead, following Westfall and Troendle (2008), we assume a family of distributions for

$$(Y_1, Y_2) = (Y_{11}, \dots, Y_{1n}, Y_{21}, \dots, Y_{2n})$$

with minimal assumptions. To introduce the assumptions, for any  $I \subseteq \{1, \dots, m\}$ , let  $Y_{ij}^I$  denote the subvector of  $Y_{ij}$  with elements in  $I$  and

$$(Y_1^I, Y_2^I) = (Y_{11}^I, \dots, Y_{1n}^I, Y_{21}^I, \dots, Y_{2n}^I).$$

The null hypotheses tested will be

$$H_i : \text{the distribution of } (Y_1^{\{i\}}, Y_2^{\{i\}}) \text{ is exchangeable.}$$

The null hypothesis says that the treatment difference in the two groups has no effect whatsoever on the  $i$ th variable. This is a natural null hypothesis for binary and nominal data. However, with interval data, if there is interest only in differences in means and not standard deviations, the permutation test may not be appropriate.

Since intersections are required for closed testing, and since multivariate permutation tests will be used for the intersections, we need to make only the

following assumption about the model. Assume that, if for  $I, J \subseteq \{1, \dots, m\}$  the distribution of  $(Y_1^I, Y_2^I)$  is exchangeable in its  $2n$  elements and the distribution of  $(Y_1^J, Y_2^J)$  is exchangeable in its  $2n$  elements, then the distribution of  $(Y_1^{I \cup J}, Y_2^{I \cup J})$  is also exchangeable in its  $2n$  elements. In particular, the assumption implies that

$$\cap_{i \in I} H_i = H_I : \text{ the distribution of } (Y_1^I, Y_2^I) \text{ is exchangeable,}$$

for any subset  $I$ .

Like all assumptions, this one may be questionable, but it should be noted that

- the model is substantially more general than the multivariate location-shift model, which is a special case of this model,
- it is perhaps not unrealistic to assume, e.g., that if there is no difference in treatment effect for each of variables  $\{1, 2\}$ , then the joint distribution of  $(Y_1^{\{1,2\}}, Y_2^{\{1,2\}})$  is exchangeable,
- unlike most statistical models, no assumption of independence is needed.

To define a closed testing procedure for this two-sample problem, define a test statistic  $t_i = t_i(Y_1^{\{i\}}, Y_2^{\{i\}})$  for each variable, with larger values suggesting non-exchangeability, and test each intersection hypothesis  $H_I$  using the maximum test statistic  $t_{\max}(I) = \max_{i \in I} t_i$ . The test statistics can be quite general, and are often defined in terms of the  $p$ -values themselves. The exact permutation  $p$ -value for the test of  $H_I$  is  $p_I$  which is defined as the proportion of the  $(2n)!$  permutations of the data vectors

$$(y_{11}^I, \dots, y_{1n}^I, y_{21}^I, \dots, y_{2n}^I)$$

that yield  $\max_{i \in I} T_i^* \geq \max_{i \in I} t_i$ . Since the subset pivotality condition is satisfied and maximum test statistics are used, the shortcut methods described in Section 2.8.1 apply here as well. Complete enumeration of the  $(2n)!$  permutations is not usually feasible, so  $p_I$  is instead typically approximated as follows:

- Step 1. Generate a resampled data set  $Y_{ij}^*$ ,  $i = 1, 2$ ,  $j = 1, \dots, n$ , by sampling *without replacement* from the observed vectors  $\{y_{11}, \dots, y_{1n}, y_{21}, \dots, y_{2n}\}$ .
- Step 2. Compute the statistics  $T_i^*$  from the  $Y_{ij}^*$ .

Repeat the two steps  $B$  times and define the exact permutation  $p$ -value  $p_I$  as the proportion of the  $B$  samples where  $\max_{i \in I} T_i^* \geq \max_{i \in I} t_i$ .

In contrast to the approximate  $p$ -values for the bootstrap-based procedures in Sections 2.8.2–2.8.3, the permutation  $p$ -values are exact when the  $(2n)!$  permutations are completely enumerated. The algorithm above can approximate

**TABLE 2.8:** Two-sided raw  $p$ -values and adjusted  $p$ -values produced by the Holm and permutation-based procedures for five adverse events. The asterisk identifies the adjusted  $p$ -values that are significant at the 0.05 level.

Adverse event	Raw $p$ -value	Adjusted $p$ -value	
		Holm	Permutation
AE1	0.0008	0.0210*	0.0021*
AE8	0.0293	0.7912	0.1340
AE6	0.0601	1.0000	0.2615
AE5	0.2213	1.0000	0.6279
AE10	0.2484	1.0000	0.9276

$p_I$  with arbitrary precision. Thus, the permutation-based procedure is effectively exact, incorporates relevant correlation between variables, and can be used even when  $v$  is much larger than  $2n$ , e.g., for gene expression data.

See Puri and Sen (1971) for further details on multivariate permutation tests. Further details and applications of resampling-based testing are given in Westfall and Young (1993). Resampling is also quite useful in constructing multiple testing procedures that control generalized definitions of the family-wise error rate, e.g., gFWER or FDP defined in Section 2.2.2. Van der Laan et al. (2004) gave methods to adapt procedures that control the FWER to control the gFWER or FDP by enlarging the rejection set. Korn et al. (2004) showed how more powerful procedures can be obtained in a straightforward fashion.

As an illustration, we will consider the adverse event data set provided by Westfall et al. (1999, Section 12.2). There are two groups, control and treatment, with 80 patients in each group, and  $m = 28$  adverse event variables (binary indicators) per patient. Null hypotheses are that the adverse events are exchangeable in the combined sample, tested using Fisher exact upper tailed  $p$ -values, with smaller  $p$ -values indicating more adverse events in the treatment group. Raw  $p$ -values, Holm step-down  $p$ -values and permutation-adjusted step-down  $p$ -values for the five most significant adverse events labeled AE1, AE8, AE6, AE5 and AE10 are shown in Table 2.8. The adjustment is performed using the minimum  $p$ -value method, which is identical to the method based on the maximum test statistic, where the test statistics  $q_j$ 's are defined by  $q_j = 1 - p_j$  and  $p_j$ 's are the unadjusted  $p$ -values from the permutation test.

There is a substantial benefit in using permutation-based step-down testing rather than the Holm method as is seen by comparing the smallest adjusted  $p$ -values in Table 2.8. This occurs because the permutation-based method takes advantage of the discreteness of the data. Permutational methods implicitly exploit sparseness. If a particular variable has a permutational distribution

that does not contain any large test statistic values, it is effectively removed from the maximum in the calculation of maximum test statistics (Westfall and Wolfinger, 1997).

---

## 2.9 Software implementation

This section describes software implementation of multiple testing procedures described in this chapter with emphasis on SAS and R.

### 2.9.1 Multiple comparison procedures in SAS

PROC MULTTEST supports a host of popular  $p$ -value-based procedures described in Section 2.6 and resampling-based procedures introduced in Section 2.8. This includes the calculation of adjusted  $p$ -values for the Bonferroni, Holm, Hochberg and Hommel procedures as well as simultaneous confidence intervals and corresponding adjusted  $p$ -values for resampling-based procedures.

$P$ -value-based and parametric procedures are available in other SAS procedures. PROC GLM and PROC MIXED support adjusted  $p$ -values and simultaneous confidence intervals for the Bonferroni and single-step Dunnett procedures (covariate-adjusted if needed) in linear and mixed-linear models. Further, PROC GLIMMIX completely subsumes PROC MIXED and PROC GLM and enables the user to implement the  $p$ -value-based and parametric procedures supported by these two procedures as well as other multiple testing procedures, e.g., it supports the extended Shaffer-Royen analysis (Section 2.7.3). In cases where the exact distributions are unavailable, these SAS procedures use simulation (often with control-variate variance reduction to improve accuracy) to obtain critical values and adjusted  $p$ -values. Monte Carlo errors can be made negligible by specifying simulation sizes in the millions or even billions.

The following is a list of SAS programs that were used to perform multiplicity adjustments in the examples included in this chapter. The programs can be downloaded from the book's Web site (<http://www.multxpert.com>).

- Program 2.1 computes adjusted  $p$ -values for the Bonferroni, Holm, fixed-sequence, fallback, Hommel and Hochberg procedures in the dose-finding trial example (Section 2.6.10).
- Program 2.2 calculates lower limits of one-sided simultaneous confidence intervals for the Bonferroni, Holm, fixed-sequence and fallback procedures in the dose-finding trial example (Section 2.6.11).
- Program 2.3 computes adjusted  $p$ -values for the single-step and step-

down Dunnett procedures in the dose-finding trial example (Section 2.7.4).

- Program 2.4 derives lower limits of one-sided simultaneous confidence intervals for the single-step and step-down Dunnett procedures in the dose-finding trial example (Section 2.7.4).
- Program 2.5 implements the resampling-based procedures discussed in Section 2.8.

### 2.9.2 Multiple comparison procedures in R

R is a language and environment for statistical computing and graphics (Ihaka and Gentleman, 1996). It provides a wide variety of statistical and graphical techniques, and is highly extensible. The latest version of R is available at the Comprehensive R Archive Network (CRAN), which can be accessed from

<http://www.r-project.org/>

In this section we illustrate briefly the use of the *multcomp* package, which provides a variety of multiple comparison procedures for the linear and other (semi-)parametric models described in Section 2.7.5. The most recent version of the *multcomp* package is available at CRAN under the contributed packages section. For a detailed discussion of multiple comparison procedures in R we refer to Bretz, Hothorn and Westfall (2010).

We consider Scenario 1 in the dose-finding trial example from Section 2.5.3 to illustrate some of the capabilities of the *multcomp* package. We first load the *multcomp* package with the command

```
R> library(multcomp)
```

Suppose that we have a data frame `data` containing the individual observations, where the variables `resp` and `dose` denote the response and the dose group, respectively. To analyse the data, we fit a standard analysis-of-variance model with the `aov` function,

```
R> data.aov <- aov(resp ~ dose, data = data)
```

Assume that we want to perform the single-step Dunnett procedure from Section 2.7.1. The `glht` function from *multcomp* takes the fitted response model `data.aov` to perform the multiple comparisons. To be more precise, we can call

```
R> data.mc <- glht(data.aov, linfct = mcp(dose = "Dunnett"),
+   alternative = "less")
```

In this statement, we used the `mcp` function for the `linfct` argument to specify the comparison type we are interested in. Since we are interested in the Dunnett procedure, we pass over the argument `Dunnett`. Other pre-defined comparison types are also available in `multcomp`; alternatively, the constant matrix  $C$  introduced in Section 2.7.5 can be specified manually, to exactly determine the experimental questions of interest. Note that this is a one-sided testing problem and we are interested in showing an increase in the mean HDL cholesterol level; therefore we pass the `alternative = "less"` argument to `glht`.

A detailed summary of the results is available from the `summary` method associated with the `glht` function:

```
R> summary(data.mc)
      Simultaneous Tests for General Linear Hypotheses
```

*Multiple Comparisons of Means: Dunnett Contrasts*

```
Fit: aov(formula = resp ~ dose, data = data)
```

*Linear Hypotheses:*

	Estimate	Std. Error	t value	p value
1 - 0 <= 0	2.899	1.445	2.006	0.0714 .
2 - 0 <= 0	3.140	1.445	2.173	0.0493 *
3 - 0 <= 0	3.561	1.445	2.465	0.0242 *
4 - 0 <= 0	3.813	1.445	2.639	0.0151 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
(Adjusted p values reported -- single-step method)

The output shows the observed mean differences, standard errors and  $t$  statistics for Scenario 1 in [Table 2.1](#). In the last column, entitled `p value`, the adjusted  $p$ -values for the single-step Dunnett procedure are reported, which coincide with the values reported for Scenario 1 in [Table 2.4](#).

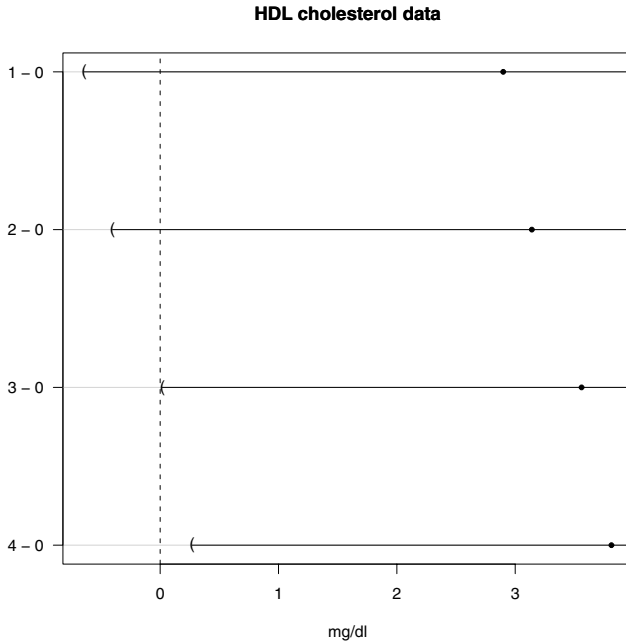
In addition, we can compute one-sided 97.5% simultaneous confidence intervals by using the `confint` method associated with the `glht` function:

```
R> data.ci <- confint(data.mc, level = 0.975)
R> data.ci
      Simultaneous Confidence Intervals
```

*Multiple Comparisons of Means: Dunnett Contrasts*

```
Fit: aov(formula = resp ~ dose, data = data)
```

Estimated Quantile = 2.4513  
97.5% family-wise confidence level



**FIGURE 2.6:** One-sided 97.5% simultaneous confidence intervals for the dose-finding trial example (Section 2.5.3, Scenario 1).

*Linear Hypotheses:*

	Estimate	<i>lwr</i>	<i>upr</i>
1 - 0 <= 0	2.89870	-0.64324	<i>Inf</i>
2 - 0 <= 0	3.14026	-0.40168	<i>Inf</i>
3 - 0 <= 0	3.56104	0.01909	<i>Inf</i>
4 - 0 <= 0	3.81299	0.27104	<i>Inf</i>

The simultaneous lower limits match with the values displayed in [Table 2.5](#). We can also display the confidence intervals graphically with

```
R> plot(data.ci, main = "HDL cholesterol data", xlab = "mg/dl")
```

see Figure 2.6 for the resulting plot.

So far we have illustrated only the single-step Dunnett procedure accounting for the correlation among the test statistics. As described in Section 2.7.2, the step-down Dunnett procedure is uniformly more powerful than the single-step Dunnett procedure. Using *multcomp*, we can perform the step-down Dunnett procedure by calling

```
R> summary(data.mc, test = adjusted(type = "free"))
      Simultaneous Tests for General Linear Hypotheses
```

Multiple Comparisons of Means: Dunnett Contrasts

```
Fit: aov(formula = resp ~ dose, data = data)
```

Linear Hypotheses:

	Estimate	Std. Error	t value	p value
1 - 0 <= 0	2.899	1.445	2.006	0.0280 *
2 - 0 <= 0	3.140	1.445	2.173	0.0280 *
3 - 0 <= 0	3.561	1.445	2.465	0.0190 *
4 - 0 <= 0	3.813	1.445	2.639	0.0152 *

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
(Adjusted p values reported -- free method)

The option `type = "free"` leads to a step-down procedure under the free combination condition, which incorporates correlations. In this example we used the Dunnett contrasts, and the results from the previous call coincide with the values reported for Scenario 1 in [Table 2.4](#). If the hypotheses are restricted, truncated closed testing procedures based on Westfall (1997) and Westfall and Tobias (2007) can be performed with the `type = "Westfall"` option; see also [Section 2.7.3](#). In combination with the parametric procedures described in [Section 2.7.5](#), the *multcomp* package thus provides powerful step-wise multiple testing procedures for a large class of parametric models, including generalized linear models, mixed models, and survival models; see Bretz, Hothorn and Westfall (2010) for further details.

The *multcomp* package also implements some of the *p*-value-based multiple comparison procedures described in [Section 2.6](#). To be more precise, *multcomp* provides an interface to the multiplicity adjustments implemented by the `p.adjust` function from the `stats` package. Given a set of (raw) *p*-values, the `p.adjust` function provides the resulting adjusted *p*-values using one of several methods, including the Bonferroni, Holm, Hochberg and Hommel procedures. In order to perform, for example, the Bonferroni procedure, one can call `summary(data.mc, test = adjusted(type = "bonferroni"))`.

Acknowledgements

James Troendle’s research was supported in part by the Intramural Research Program of the National Institute of Child Health and Human Development. Ajit C. Tamhane’s research was supported by grants from the National Heart, Lung and Blood Institute.



# Chapter 3

---

## *Multiple Testing in Dose-Response Problems*

**Frank Bretz**

*Novartis*

**Ajit C. Tamhane**

*Northwestern University*

**José Pinheiro**

*Novartis*

---

### **3.1 Introduction**

Dose-response studies play an important role in many applications. Particularly in pharmaceutical drug development the assessment of a dose-related effect and the subsequent estimation of a target dose are of central importance. Clinical dose-finding studies are designed to investigate the effect of a given drug at several doses, possibly in comparison with a placebo group or an active control. Current clinical practice typically aims at two major goals when performing a dose-finding study (Ruberg, 1995a, 1995b; Bretz et al., 2008):

- First, one is interested in assessing an overall dose-response effect. The aim is thus to ensure that, by changing the dose, the response (efficacy and/or safety) changes as well.
- Once such an overall trend has been established, the second aim is to estimate a target dose of interest; this could be, for example, the minimum effective dose; i.e., the smallest dose associated with a statistically significant and clinically relevant effect.

The importance of clinical dose-finding studies is also reflected by the early publication of the ICH E4 guideline, which is the primary source of regulatory guidance on dose response studies (ICH, 1994). The ICH E4 guideline emphasizes that “Dose response assessment should be an integral part of drug

development.” It further mentions specifically the “... smallest dose with a discernible useful effect” – which is typically interpreted as a requirement to estimate the minimum effective dose – and “... a maximum dose beyond which no further beneficial effect is seen,” also known as the maximum effective dose (Ting, 2003). Similarly, there are ongoing activities in the pharmaceutical industry to address poor performance of standard dose selection methods resulting from incorrect or incomplete knowledge of the dose response relationship for both efficacy and safety (Bornkamp et al., 2007). In light of the recent discussions and activities, this chapter aims at reviewing some of the key statistical methodologies used in dose-finding trials. For a general reading about dose finding in drug development we refer to the edited books by Ting (2006), Chevret (2006), Krishna (2006) and the references therein.

The analysis of dose-finding studies can be classified into two major strategies: modeling techniques (Pinheiro, Bretz and Branson, 2006; Bates and Watts, 1988) and multiple test procedures (Hochberg and Tamhane, 1987; Hsu, 1996). In this chapter we review multiple test procedures, which regard the dose as a qualitative factor and make very few, if any, assumptions about the underlying dose-response model. Multiple test procedures can either be used for detecting an overall dose-related effect by means of trend tests or for estimation of target doses by stepwise testing strategies, while preserving the familywise error rate (FWER) at a prespecified level  $\alpha$ . Such procedures are relatively robust to the underlying dose-response shape, but they are not designed for extrapolation of information beyond the observed dose levels. Inference is thus confined to the dose levels under investigation. Hybrid approaches have recently been proposed that combine multiple testing with modeling and thus overcome some of the inherent restrictions of multiple comparison procedures (Bretz et al., 2005). This class of methods aims, initially, at detecting a dose response signal using powerful trend tests. If such a signal is detected, the procedures continue estimating the target dose(s) of interest using modeling techniques.

This chapter is organized as follows. In Section 3.2 we review trend tests for the detection of a dose-response signal. In Section 3.3 we focus on the description of commonly used multiple comparison procedures tailored to the estimation of target doses, such as the minimum effective dose and the maximum safe dose. In Section 3.4 we discuss related power and sample size calculations. In Section 3.5 we introduce hybrid approaches, which overcome some of the inherent restrictions of multiple testing procedures by combining multiple testing with modeling. Several examples are included throughout this chapter to illustrate the various methods.

## 3.2 Dose-response trend tests

In this section we review common trend tests for the detection of a dose-response signal. In Section 3.2.1 we introduce basic notation. We then describe contrast tests both in the one-way layout (Section 3.2.2) and in the general linear model (Section 3.2.3). In Section 3.2.4 we discuss a numerical example to illustrate the main results of this section.

### 3.2.1 Notation

To simplify the description, we start by considering the one-way design. Extensions to general linear models are given later. Let

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad (3.1)$$

denote the observation on the  $j$ th subject at dose level  $i$ ,  $j = 1, \dots, n_i$ ,  $i = 0, \dots, m$ , where  $m \geq 2$ . Here, the  $\mu_i$ 's denote the mean treatment effects and the  $\varepsilon_{ij}$ 's denote the independent normally distributed random errors with mean 0 and variance  $\sigma^2$ . The index  $i = 0$  denotes a zero dose control or a placebo group. Further, let  $\bar{y}_i$  and  $\bar{y}$  denote the arithmetic means in group  $i = 0, \dots, m$ , and overall mean, respectively, i.e.,

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}, \quad \bar{y} = \frac{1}{n} \sum_{i=0}^m n_i \bar{y}_i.$$

Also, let  $s^2$  denote the pooled variance estimate with  $\nu = N - (m + 1)$  degrees of freedom, where  $N = \sum_{i=0}^m n_i$  is the total number of patients, i.e.,

$$s^2 = \frac{1}{\nu} \sum_{i=0}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

We are interested in testing the null hypothesis

$$H : \mu_0 = \dots = \mu_m \quad (3.2)$$

of no differential effect among the  $m$  dose groups against the restricted alternative

$$K : \mu_0 \leq \dots \leq \mu_m \text{ with } \mu_0 < \mu_m. \quad (3.3)$$

Note that the formulation of the alternative hypothesis  $K$  explicitly assumes the treatment means  $\mu_i$  to be monotonically ordered (without loss of generality, we will assume monotonically increasing profiles, if not stated otherwise). This is a strong assumption and may be violated in practice, for example, when a downturn in effect occurs at larger dose levels (so-called umbrella-shaped dose-response curves). Nevertheless, if the monotonicity condition 3.3

holds, powerful tests can be derived. Test procedures, which do not rely on this assumption, are by construction less powerful if monotonicity in fact holds. However, caution is advisable if one is not sure whether the dose-response shape is truly monotonic. Bauer (1997) showed that already small departures from the assumed monotonicity may lead to invalid results if trend tests, such as those presented below, are used. They are then not useful in the sense that they do not control the probability of incorrectly declaring a dose to be effective when in fact it is not. Irrespective of which trend test is going to be conducted, the decision for its use should always take into account the plausibility of the monotonicity assumption, and the context of the application should be analyzed carefully before looking at the data. Generalizations to situations, in which a possible downturn at high doses cannot be excluded a priori, are handled by Simpson and Margolin (1990), Pan and Wolfe (1996) and Bretz and Hothorn (2001). In what follows, we review standard trend tests under the monotonicity assumption 3.3. In Section 3.5 we will investigate alternative approaches, which also allow for the possibility of an umbrella shape.

### 3.2.2 Contrast tests

Dose-response trend tests have been investigated since many decades. Well-known trend tests developed under the classical ANOVA assumptions 3.1 include the likelihood ratio test for homogeneity of normal means under the total order restriction (Bartholomew, 1961), the modified two-sample  $t$ -test of Williams (1971) and its modified version investigated by Marcus (1976). Common to these and other approaches, however, are their restrictive assumptions. Critical values are often available only for the balanced case and the methods cannot be used in the presence of covariates. In the following, we thus concentrate on contrast tests, as first introduced by Abelson and Tukey (1963) and Schaafsma and Smid (1966) in the context of dose-response testing. These are powerful methods to detect dose-response and can be applied to a variety of different statistical models, including general linear models allowing for covariates and/or factorial treatment structures (Hothorn et al., 2008).

Let  $c_0, \dots, c_m$  denote fixed constants such that  $\sum_{i=0}^m c_i = 0$  ( $c_0, \dots, c_m$  are known as contrast coefficients). A single contrast test is based on the test statistic

$$t = \frac{\sum_{i=0}^m c_i \bar{y}_i}{s \sqrt{\sum_{i=0}^m \frac{c_i^2}{n_i}}}.$$

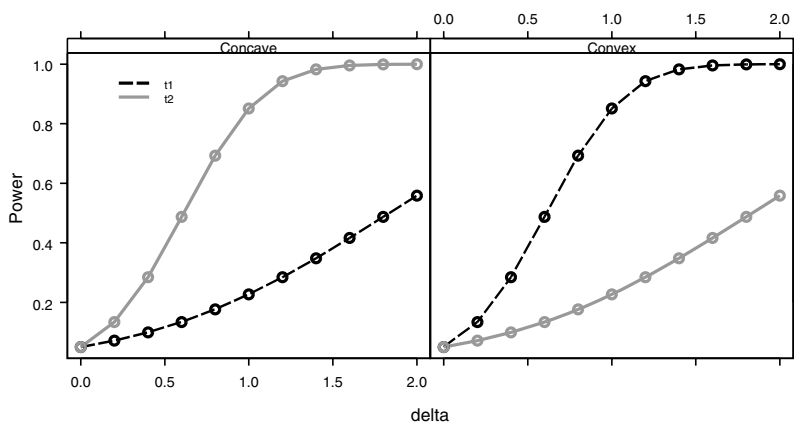
By construction, the test statistic  $t$  follows a central  $t$ -distribution with  $\nu$  degrees of freedom under the null hypothesis  $H$  from 3.2. When  $H$  is not true,

$t$  follows a non-central  $t$ -distribution with non-centrality parameter

$$\tau = \frac{\sum_{i=0}^m c_i \mu_i}{\sigma \sqrt{\sum_{i=0}^m \frac{c_i^2}{n_i}}}.$$

Numerous proposals for the choice of the contrast coefficients have been made; we refer to Tamhane et al. (1996) for some examples. The optimal choice depends on the unknown true mean values  $\mu_0, \dots, \mu_m$ , specifically, on the correlation between the  $c_i$ 's and the  $\mu_i$ 's.

Figure 3.1 illustrates how the choice of contrast coefficients impacts the power of a single contrast test. Suppose that we compare  $m = 3$  dose levels of a new drug with placebo. We investigate the following two sets of contrast coefficients. Suppose the first contrast test  $t_1$  uses  $c_0 = c_1 = c_2 = -1$ ,  $c_3 = 3$  and the second contrast test  $t_2$  uses  $c_0 = -3$ ,  $c_1 = c_2 = c_3 = 1$ . We compute the power at the two configurations  $\mu_0 = 0$ ,  $\mu_1 = \mu_2 = \mu_3 = \delta$  (concave shape) and  $\mu_0 = \mu_1 = \mu_2 = 0$ ,  $\mu_3 = \delta$  (convex shape) for varying values of the shift parameter  $\delta \in [0, 2]$ . The results from Figure 3.1 indicate that the power of single contrast tests depends substantially on the extent to which the contrast coefficients imitate the pattern of the  $\mu_i$ 's. In extreme cases the potential loss in power can be as high as 60%.



**FIGURE 3.1:** Power values for  $t_1$  (dotted line) and  $t_2$  (solid line) with  $n_i = 10$ ,  $\sigma = 1$ ,  $\alpha = 0.05$  for the concave and convex mean configurations.

Multiple contrast tests have been introduced to achieve more robustness with respect to misspecification of the contrast coefficients. The key

idea is to identify “... a set of vectors that are ‘strategically’ located within the alternative region” (Mukerjee, 1987). Consider  $q \geq 2$  contrast vectors,  $\mathbf{c}_j = (c_{j0}, \dots, c_{jm})'$ ,  $j = 1, \dots, q$ . Let  $t_1, \dots, t_q$  denote the corresponding test statistics. Then the multiple contrast test uses the test statistic

$$t_{\max} = \max \{t_1, \dots, t_q\}. \quad (3.4)$$

Assuming the normal model 3.1, it can be shown that under the null hypothesis  $H$  the test statistics  $t_1, \dots, t_q$  are jointly multivariate  $t$ -distributed with  $\nu$  degrees of freedom and correlation matrix  $\mathbf{R} = (\rho_{ij})$ , where

$$\rho_{ij} = \frac{\sum_{\ell=0}^m \frac{c_{i\ell} c_{j\ell}}{n_\ell}}{\sqrt{\left(\sum_{\ell=0}^m \frac{c_{i\ell}^2}{n_\ell}\right) \left(\sum_{\ell=0}^m \frac{c_{j\ell}^2}{n_\ell}\right)}}, \quad 0 \leq i, j \leq m. \quad (3.5)$$

Let  $t_\alpha(q, \nu, \mathbf{R})$  denote the upper  $\alpha$  equicoordinate critical point of this distribution. Then the multiple contrast test rejects  $H$  if  $t_{\max} > t_\alpha(q, \nu, \mathbf{R})$ . We refer to Kotz and Nadarajah (2004) for a general overview of the multivariate  $t$ -distribution and to Genz and Bretz (2002, 2009) for details on computing multivariate  $t$  probabilities.

The class of multiple contrast tests is very broad and includes many common multiple test procedures. Bretz et al. (2001) provided a list of several multiple comparison procedures (not necessarily restricted to trend tests), which can be formulated as multiple contrast tests, including the many-to-one comparisons (Dunnett, 1955), all-pair comparisons (Tukey, 1953) and multiple comparisons with the best (Hsu, 1984). Moreover, many common trend tests can be formulated as multiple contrast tests, including the Williams test and its modification mentioned earlier in this section (Bretz, 2006). In Section 3.5 we describe an approach to derive optimal (in a certain sense) contrast coefficients for a given candidate set of potential dose-response models.

### 3.2.3 Multiple contrast tests for general linear models

We now extend the description of multiple contrast tests to the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (3.6)$$

where  $\mathbf{y}$  is an  $N \times 1$  observation vector,  $\mathbf{X}$  is a fixed and known  $N \times p$  design matrix (which allows for covariates and/or factorial treatment structures),  $\boldsymbol{\beta}$  is a fixed and unknown  $p \times 1$  parameter vector and  $\boldsymbol{\varepsilon}$  is a random, unobservable  $N \times 1$  error vector, with elements  $\varepsilon_i$  being independent normally distributed random errors with mean 0 and variance  $\sigma^2$ . Note that in the context of the general linear model 3.6, we use a different notation for the description of the methods in order to conform with the conventions from the literature. The one-way design from equation 3.1 specified without an intercept term is obtained

as a special case from 3.6 by setting  $p = m + 1$  and  $\beta_i = \mu_i$ ,  $i = 0, \dots, m$ , where  $m$  denotes the number of non-zero dose levels under investigation.

We assume the usual least square estimates

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} \text{ and } s^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{\nu},$$

where  $\nu = N - \text{rank}(\mathbf{X})$  and  $(\mathbf{X}'\mathbf{X})^{-}$  denotes a generalized inverse of  $\mathbf{X}'\mathbf{X}$ . We are interested in the pivotal quantities

$$t_j = \frac{\mathbf{c}_j'\hat{\beta}}{s\sqrt{\mathbf{c}_j'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}_j}}, \quad j = 1, \dots, q, \quad (3.7)$$

where  $q$  is the number of single contrast tests. The elements of the  $p \times 1$  vectors  $\mathbf{c}_j$  are similarly interpreted as in the case of single contrast tests, where now the weights are not only assigned to the treatment levels but also to the other elements of  $\beta$ , such as covariates, for example. The joint distribution of  $t_1, \dots, t_q$  is again multivariate  $t$  with  $\nu$  degrees of freedom and correlation matrix  $\mathbf{R} = \mathbf{D}\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}\mathbf{D}$ , where  $\mathbf{C}_{m \times q} = (\mathbf{c}_1, \dots, \mathbf{c}_q)$  and  $\mathbf{D} = \text{diag}(\mathbf{c}_j'(\mathbf{X}'\mathbf{X})^{-}\mathbf{c}_j)^{-1/2}$ . In the asymptotic case ( $\nu \rightarrow \infty$ ) or if  $\sigma$  is known, the corresponding multivariate normal distribution holds in the limit. Numerical integration methods to compute the associated probabilities are described by Genz and Bretz (2002, 2009) and implemented in the R package `mvtnorm`. In the special case of a one-way design, the entries of  $\mathbf{R}$  reduce to the correlation coefficients given by equation (3.5).

### 3.2.4 Thalidomide study example

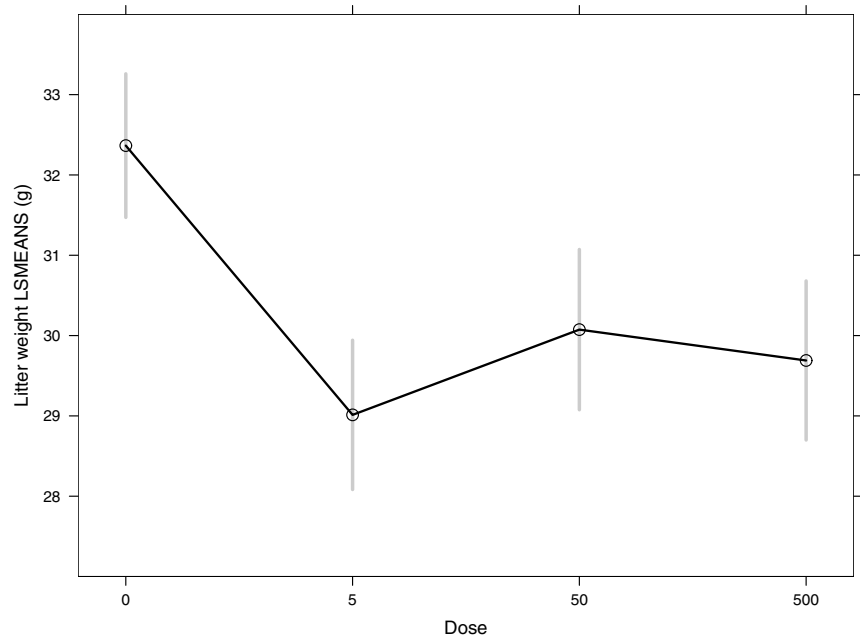
We use a dose-response study involving two covariates to illustrate some of the methods described in the previous subsections. The data of the Thalidomide study described in Westfall and Young (1993, Example 3.4) are available at the companion Web site for Westfall et al. (1999):

<http://ftp.sas.com/samples/A56648>

Figure 3.2 displays the summary statistics in the thalidomide study. Four dose levels (0, 5, 50 and 500 units of a study compound) were administered to pregnant mice. Their litters were evaluated for defects and weights. Westfall and Young (1993) proposed to reduce the multiplicity problem by combining the weekly weights into a single variable of interest. Thus, the primary response variable is the average post-birth weight in the entire litter for the first three weeks of the study. Since it is well known that the litter weight depends on gestation time and number of siblings in the litter, these two variables were included as covariates in the analysis. Consequently, we assume the normal linear model

$$E(y_{ij}) = \beta_0 + \beta_i + \beta_5 z_{1ij} + \beta_6 z_{2ij}, \quad (3.8)$$

where  $y_{ij}$  denotes the weight of the  $j$ th litter at dose level  $i$ ,  $z_1$  and  $z_2$  denote the two covariates and  $\beta_i$  denotes the effect of the dose level  $i$  after adjusting for the effects of the covariates,  $i = 1, \dots, 4$ .



**FIGURE 3.2:** Summary plot of the litter weight data in the thalidomide study.

The experimental question is whether we can assess a statistically significant decrease in average post-birth weight with increasing doses of Thalidomide. Note that because of using an ANCOVA model with unequal group sample sizes, many standard trend tests, such as the likelihood ratio test, the Williams test or its modification, cannot be applied. In the following we apply a multiple contrast test taking the covariate information into account. More specifically, we use contrast versions of the Williams test and its modification to assess dose-response. The results from this example also motivate some of the developments in Section 3.5.

Consider model 3.8, where the parameter vector  $\beta$  consists of  $p = 7$  elements  $\beta_i$ . We are interested in testing the null hypothesis of no treatment effect,

$$H : \beta_1 = \dots = \beta_4,$$



against the ordered alternative

$$K : \beta_1 \geq \dots \geq \beta_4 \text{ with } \beta_1 > \beta_4.$$

We conclude in favor of  $K$  if the related trend test is significant. Note that in this example we are interested in detecting a reduction of weight with increasing doses. Therefore, we test for a decreasing trend in this example (as opposed to the previous sections, where we assumed an increasing trend).

We first consider the Williams test extended to the framework of multiple contrast tests, as described in Bretz (2006). The group sample sizes are  $n_i = 20, 19, 18, 17$  for  $i = 1, 2, 3, 4$ , respectively. With these sample sizes, the contrast coefficients are given by

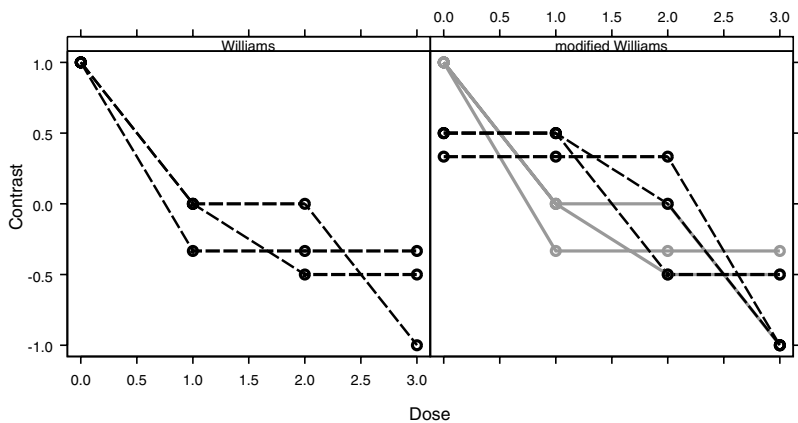
$$\mathbf{C}'_W = \begin{pmatrix} 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -0.5143 & -0.4857 & 0 & 0 \\ 0 & 1 & -0.3519 & -0.3333 & -0.3148 & 0 & 0 \end{pmatrix}.$$

Figure 3.3 displays the contrast coefficients for  $\beta_1, \dots, \beta_4$ . Each single contrast test consists of a comparison between the control group and a weighted average over the last  $\ell = 1, \dots, 3$  treatment groups, respectively. To illustrate this, consider as an example the second contrast. The control group is compared with the weighted average of the two highest dose levels. The contrast coefficient for the control group is 1. The contrast coefficients for the largest and second largest dose levels are  $-17/(17 + 18) = -0.4857$  and  $-18/(17 + 18) = -0.5143$ , respectively; here, 17 and 18 are the respective group sample sizes. The lowest dose level is not included in this comparison and is therefore assigned the contrast coefficient 0. Note that the weights for the intercept  $\beta_0$  and the two covariate parameters  $\beta_5$  and  $\beta_6$  are set to 0, so that  $\mathbf{C}'_W \boldsymbol{\beta}$  involves only the comparison of the parameters of interest. As shown by Bretz (2006), this representation of  $\mathbf{C}_W$  ensures that the same type of comparison is performed as for the original Williams test while accounting for the covariate information through the computation of the least square estimates  $\hat{\boldsymbol{\beta}}$  and  $s$ . Note that the sign of the contrast matrix is chosen such that the individual contrast test statistics remain positive for decreasing weights.

The modified Williams test (Marcus, 1976) can be extended similarly and the corresponding contrast coefficients are given by

$$\mathbf{C}'_{MW} = \begin{pmatrix} 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -0.5143 & -0.4857 & 0 & 0 \\ 0 & 1 & -0.3519 & -0.3333 & -0.3148 & 0 & 0 \\ 0 & 0.5128 & 0.4872 & -0.5143 & -0.4857 & 0 & 0 \\ 0 & 0.5128 & 0.4872 & 0 & -1 & 0 & 0 \\ 0 & 0.3509 & 0.3333 & 0.3158 & -1 & 0 & 0 \end{pmatrix}.$$

Figure 3.3 also plots the contrast coefficients for this contrast version of the modified Williams test. Note that  $\mathbf{C}_W$  is a subset of  $\mathbf{C}_{MW}$  in the sense that the contrast vectors for the former test are all contained in the contrast



**FIGURE 3.3:** Graphical representation of contrast coefficients for the Williams and modified Williams tests (balanced case).

matrix of the latter test (the first three rows). These weights are particularly suitable for testing concave dose-response shapes, as the higher dose groups are pooled and compared with the zero dose group. Rows five and six of  $\mathbf{C}_{MW}$  are appropriate for detecting convex shapes, as they average the lower treatments. The fourth contrast is particularly powerful for linear or almost linear relationships. To illustrate the computation of the coefficients for the fourth contrast, we note that the weighted average of the two higher dose levels is compared with the weighted average of the remaining treatments. With the given group sample sizes, we therefore obtain  $20/(20 + 19) = 0.5128$ ,  $19/(20 + 19) = 0.4872$ ,  $18/(18 + 17) = 0.5143$ , and  $17/(18 + 17) = 0.4857$ . See Bretz (2006) for the general expressions of the contrast coefficients.

Table 3.1 reports the results of the tests. The individual contrast test statistics were obtained from 3.7 based on model 3.8. Unadjusted  $p$ -values were calculated from the marginal univariate  $t$  distribution. Adjusted  $p$ -values for the individual comparisons were obtained from the multivariate  $t$ -distribution, based on either  $\mathbf{C}_W$  or  $\mathbf{C}_{MW}$ . For comparison purposes we also included the Bonferroni adjustment, which does not take the correlation between the test statistics into account and essentially multiplies the individual  $p$ -values by 3 (Williams contrast test) or 6 (modified Williams contrast test). The adjusted  $p$ -value for the Williams contrast test taking the maximum over the individual contrast test statistics is the minimum of the  $q = 3$   $p$ -values shown in Table 3.1. In our example,  $p = 0.0095$  and we conclude for a significant dose-response signal at the 5% significance level. For the modified Williams contrast test we obtain  $p = 0.0155$ . The computations can be done using standard software,

such as the `multcomp` package in R (Bretz et al., 2010) or the `%sim*` macros in SAS from Westfall et al. (1999).










Because the maximum contrast test based on  $t_{\max}$  is a union-intersection test, it is consonant and we know which individual contrast test is significant (Gabriel, 1969). Thus, additional information about the true dose-response shape is given, as long as individual statements are set into correspondence with the dose-response shapes (Tukey et al., 1985; Stewart and Ruberg, 2000). In case of standard trend tests, a significant  $p$ -value yields no further conclusions. Consider the dose-response shapes in the left column of Table 3.1, which essentially associate the shapes from Figure 3.3 with the (adjusted)  $p$ -values for the individual contrast test statistics. In this numerical example, the individual tests, which pool the higher doses, have considerably smaller  $p$ -values than the tests, which pool the lower doses. Thus, conclusions in view of model selection would suggest a significant convex dose-response shape over the dose range under investigation. We will further develop and formalize these ideas in Section 3.5.

---

### 3.3 Target dose estimation using multiple hypothesis testing

This section discusses the estimation of appropriately defined target doses in dose-response studies with efficacy and safety endpoints. From an efficacy perspective, the target dose of primary interest is the *minimum effective dose* (MINED), which is the lowest dose whose effect exceeds that of the control by a specified threshold. If the efficacy response is non-monotone, e.g., umbrella-shaped, then the *maximum effective dose* (MAXED) is also of interest. For safety or toxicity endpoints, the target dose of primary interest is the *maximum safe dose* (MAXSD). This dose is defined as the highest dose whose adverse effect or toxicity is less than that of the control by a specified threshold. The interval between the MINED and MAXSD is called the *therapeutic window*, assuming it is nonempty. We will begin with the problem of estimating the MINED in Section 3.3.1 and then the problem of estimating the MAXSD in Section 3.3.2 will be discussed. Tamhane and Logan (2006) gave a review of procedures for estimation of these quantities. Note that other multiple test strategies to estimate alternative target doses of interest are available. For example, Strassburger et al. (2007) used the partitioning principle to estimate the MAXED. We refer to the original publications for details.

**TABLE 3.1:** Results for the litter weight data example (corrected from Bretz, 2006).

Contrast coefficients	Value of test statistic	Unadjusted <i>p</i> -value	Adjusted <i>p</i> -value	
			Bonferroni	Multivariate <i>t</i>
<i>Williams contrast test</i>				
	2.0050	0.0245	0.0735	0.0438
	2.2072	0.0153	0.0459	0.0283
	2.6616	0.0048	0.0144	0.0095
<i>Modified Williams contrast test</i>				
	2.0050	0.0245	0.1470	0.0679
	2.2072	0.0153	0.0918	0.0446
	2.6616	0.0048	0.0288	0.0155
	0.8963	0.1866	1	0.3778
	0.8900	0.1883	1	0.3805
	0.7350	0.2324	1	0.4469

### 3.3.1 Estimation of the minimum effective dose

Consider the primary efficacy endpoint in a dose-response study and assume that higher values of this endpoint indicate an improvement in the condition of interest.

Two measures of difference used to quantify the efficacy of a dose are defined below.

- The difference measure  $\delta_i = \mu_i - \mu_0$ ,  $1 \leq i \leq m$ . Dose  $i$  is regarded effective if the difference is greater than a specified value  $\delta > 0$ .
- The ratio measure  $\lambda_i = \mu_i/\mu_0$ ,  $1 \leq i \leq m$  (the use of the ratio measure assumes that  $\mu_0 > 0$ ). Dose  $i$  is declared effective if the ratio exceeds a specified threshold,  $\lambda > 1$ .

In the following we will focus on the ratio measure.

A common definition of MINED is

$$\text{MINED} = \min\{i : \lambda_i > \lambda\}. \quad (3.9)$$

A stronger definition is

$$\text{MINED} = \min\{i : \lambda_j > \lambda \text{ for all } i \leq j \leq m\}. \quad (3.10)$$

If the dose-response function is monotone then the two definitions are equivalent. Since dose-response functions are often non-monotone, the weaker definition 3.9 is generally employed.

Let  $\widehat{\text{MINED}}$  denote the estimated MINED. We want to control the error probability of any ineffective dose being declared effective to be no more than a specified level  $\alpha$ , i.e.,

$$P(\widehat{\text{MINED}} < \text{MINED}) \leq \alpha. \quad (3.11)$$

This is achieved by strong control of the FWER for the multiple hypotheses testing problem:

$$H_i : \mu_i \leq \lambda\mu_0 \text{ vs. } K_i : \mu_i > \lambda\mu_0, \quad 1 \leq i \leq m. \quad (3.12)$$

If MINED is defined using the definition 3.9,  $\widehat{\text{MINED}}$  is defined as the first hypothesis in the sequence which is found false, i.e.,

$$\widehat{\text{MINED}} = \min\{i : H_i \text{ is rejected}\}. \quad (3.13)$$

Similarly, when the definition 3.10 is utilized,  $\widehat{\text{MINED}}$  is defined as

$$\widehat{\text{MINED}} = \min\{i : H_j \text{ is rejected for all } i \leq j \leq m\}. \quad (3.14)$$

In this section we discuss two step-down (SD) procedures for the estimation of MINED. We consider only SD procedures based on pairwise contrasts (PC) because

- only pairwise contrasts yield procedures that control the FWER even when the dose-response is not monotone (Bauer, 1997);
- the procedures are simple to use;
- the procedures can be easily extended to non-normal data by using appropriate two-sample statistics.

The SD1PC procedure estimates the MINED according to definition 3.9, while the SD2PC procedure estimates the MINED according to definition 3.10. Under monotonicity, the two definitions are equivalent and either SD1PC or SD2PC may be used. These procedures were proposed by Tamhane, Hochberg and Dunnett (1996).

### SD1PC Procedure

The FWER for the family  $F = \{H_1, \dots, H_m\}$  can be controlled by applying the closure principle of Marcus et al. (1976) introduced in Section 2.3. Let  $H'_i = \bigcap_{j=1}^i H_j$ . Then it is easily checked that the closure of family  $F$  is given by  $\bar{F} = \{H'_i, 1 \leq i \leq m\}$  and the hypotheses  $H'_i$  are nested, i.e.,  $H'_m \subseteq \dots \subseteq H'_1$ . Note that this does not require the monotonicity of the dose response assumption. Therefore an FWER controlling step-down shortcut to the closed procedure is obtained by testing the hypotheses  $H'_m, \dots, H'_1$  in a stepwise manner, each at level  $\alpha$ , stopping and retaining the hypotheses  $H'_j$  for  $j \leq i$  if the hypothesis  $H'_i$  is retained. If a union-intersection (UI) statistic (see Section 2.3) is used for testing and if  $H'_i$  is the last rejected hypothesis then dose  $i$  is MINED.

The SD1PC procedure is the above step-down procedure that uses the pairwise  $t$ -statistics

$$t_i = \frac{\bar{y}_i - \lambda \bar{y}_0}{s \sqrt{\lambda^2/n_0 + 1/n_i}} \quad (3.15)$$

for testing  $H_1, \dots, H_m$ . The UI statistic for testing  $H'_i$  is

$$t_{i,\max} = \max_{1 \leq j \leq i} t_j.$$

Let  $T_k$  be random variables corresponding to  $t_i$ ,  $1 \leq k \leq m$ . Assume the least favorable configuration which maximizes the FWER, i.e.,  $\mu_j = \lambda \mu_0$  for all  $1 \leq j \leq k$ . Then, under  $H'_k$ , the joint distribution of  $T_1, \dots, T_k$  is a  $k$ -variate  $t$ -distribution with  $\nu = N - (m + 1)$  degrees of freedom and correlation matrix  $\mathbf{R}_k = \{\rho_{ij}\}$  with a product correlation structure, i.e.,  $\rho_{ij} = \tau_i \tau_j$ , where

$$\tau_i = \frac{\lambda}{\sqrt{\lambda^2 + r_i}}, \quad r_i = \frac{n_0}{n_i}, \quad 1 \leq i \leq k. \quad (3.16)$$

Note that in a balanced case; i.e., when  $n_1 = \dots = n_m = n$  and  $r = n_0/n$ , the

off-diagonal elements in the correlation matrix are given by  $\rho_{ij} = \lambda^2/(\lambda^2 + r)$ ,  $i \neq j$ .

Let  $t_\alpha(k, \nu, \mathbf{R}_k)$  denote the upper  $\alpha$  equicoordinate critical point of the  $k$ -variate  $t$ -distribution with correlation matrix  $\mathbf{R}_k$ . These critical points can be computed using the algorithms in Genz and Bretz (2009). The SD1PC procedure rejects  $H'_k$  if and only if  $H'_m, \dots, H'_{k+1}$  are rejected and  $t_{k, \max} > t_\alpha(k, \nu, \mathbf{R}_k)$ .

## SD2PC Procedure

If the dose-response can be assumed to be monotone then  $H'_i = H_i$ , so one can use the following simplified closed step-down procedure: Test the hypotheses  $H_m, \dots, H_1$  in a stepwise manner, each at level  $\alpha$ , stopping and retaining the hypotheses  $H_j$  for  $j \leq i$  if the hypothesis  $H_i$  is retained. This is the SD2PC procedure that uses the pairwise  $t$ -statistic 3.15 to test  $H_i$  and rejects it if  $H_j$  for  $j > i$  are rejected and  $t_i > t_\alpha(\nu)$ , where  $t_\alpha(\nu)$  is the upper  $\alpha$  critical point of the  $t$ -distribution with  $\nu$  degrees of freedom.

Under the monotonicity assumption, the SD2PC procedure is more powerful than the SD1PC procedure. However, if there is a downturn in the dose-response function at high dose levels, the SD2PC procedure can stop too soon and fail to identify the correct MINED. This is analogous to the fixed-sequence procedure that stops sooner if an earlier placed test in the sequence is not sufficiently powered (see Section 2.6).

## Simultaneous confidence intervals

Simultaneous confidence intervals for the ratios  $\lambda_1, \dots, \lambda_m$  can be constructed using the method developed by Bretz, Hothorn and Hsu (2003). Consider the pairwise test statistics defined in (3.15). Each statistic follows a  $t$  distribution with  $\nu$  degrees of freedom. The marginal  $100(1 - \alpha)\%$  lower confidence limit for  $\lambda_i$ ,  $1 \leq i \leq m$ , is found from the inequality  $t_i \leq t_\alpha(\nu)$ . The limit is given by

$$L_i = \frac{\bar{y}_0 \bar{y}_i - \sqrt{a_0 \bar{y}_i^2 + a_i \bar{y}_0^2 - a_0 a_i}}{\bar{y}_0^2 - a_0},$$

where  $a_i = t_\alpha^2(\nu) S^2 / n_i$ ,  $0 \leq i \leq m$ .

Given the marginal lower confidence limits, simultaneous lower confidence limits, denoted by  $\tilde{L}_i$ ,  $1 \leq i \leq m$ , can be obtained similar to Hsu and Berger (1999); see Section 2.6.11. The confidence limits are computed using the following sequential procedure.

**Step 1.** If  $L_m \leq \lambda$  then  $\tilde{L}_k = L_m$  and  $\tilde{L}_j$  is undefined for  $1 \leq j < m$ . Otherwise, let  $\tilde{L}_m = \lambda$  and go to Step 2.

**Step  $i = 2, \dots, m$ .** If  $L_{m-i+1} \leq \lambda$  then  $\tilde{L}_{m-i+1} = L_{m-i+1}$  and  $\tilde{L}_j$  is unde-

**TABLE 3.2:** Summary statistics for HAMD-17 decreases from baseline in the major depressive disorder trial.

	Treatment group				
	P	D1	D2	D3	D4
$n$	85	87	88	87	85
Mean	6.1	8.3	9.3	11.1	10.3
Pooled SD	6.6	6.6	6.6	6.6	6.6

**TABLE 3.3:** Test statistics in the major depressive disorder trial.

	Comparison			
	D1 vs. P	D2 vs. P	D3 vs. P	D4 vs. P
$t$ -statistic	0.881	1.783	3.396	2.665

fined for  $1 \leq j < m - i + 1$ . Otherwise, let  $\tilde{L}_{m-i+1} = \lambda$  and go to Step  $i + 1$ .

**Step  $m + 1$ .** Let  $\tilde{L}_i = \min_{1 \leq j \leq m} L_j$  for all  $1 \leq i \leq m$ .

Note that a MINED estimation procedure based on these lower confidence limits is equivalent to the SD2PC procedure.

**Estimation of MINED in a major depressive disorder trial**

Consider a Phase II placebo-controlled parallel group clinical trial of a new treatment for patients with major depressive disorder. Four dose levels were utilized in the study, 10, 20, 40 and 60 mg/day. The four dose levels will be labeled D1 through D4, respectively, and placebo will be labeled P. A total of 432 patients were randomized to the five treatment groups. The efficacy of the treatment was evaluated using the mean reduction from baseline in the 17-item Hamilton Depression Scale (HAMD-17). The summary statistics for the HAMD-17 score are displayed in Table 3.2.

A 20% improvement in the HAMD-17 score relative to placebo group will be used as a threshold for defining a clinically significant improvement, i.e.,  $\lambda = 1.2$ . Since the sample sizes were approximately equal across the treatment groups in the trial, we will let  $r_i = 1$  and  $\rho_{ij} = 1.2^2/(1.2^2 + 1) = 0.590$ . The  $t$ -statistics computed using 3.15 are shown in Table 3.3.

The SD1PC and SD2PC procedures will be applied to estimate MINED at a one-sided  $\alpha = 0.025$ . The SD1PC procedure begins with the hypothesis  $H_4$  (comparison between D4 and P). The test is performed by comparing  $t_{4,\max} = 3.396$  with the critical value  $t_{0.025}(4, 427, 0.590) = 2.427$ . The test statistic is greater than 2.427 and the SD1PC procedure proceeds to test  $H_3$ . Note that in this particular case  $t_{4,\max} = t_3 = 3.396$  and thus we can skip  $H_3$  and consider  $H_2$ . At this step  $t_{2,\max} = 1.738$  is compared with



$t_{0.025}(2, 427, 0.590) = 2.207$ . The test statistic does not exceed the critical value and therefore  $H_2$  is retained. The hypothesis  $H_1$  is retained by implication and, as a consequence,  $\widehat{\text{MINED}} = 3$ .

The SD2PC procedure also begins with  $H_4$ . It is easy to see that  $t_4 = 2.665$  and  $t_3 = 3.396$  are both greater than the critical value  $t_{0.025}(427) = 1.966$  and thus  $H_4$  and  $H_3$  are rejected. Further,  $t_2 = 1.783$  is less than 1.966 and the SD2PC procedure stops and retains both  $H_2$  and  $H_1$ . This leads to the same conclusion as above, i.e.,  $\widehat{\text{MINED}} = 3$ .

Using the method proposed by Bretz, Hothorn and Hsu (2003), the marginal 97.5% lower confidence limits for the  $\lambda$ 's are given by

$$L_1 = 1.031, L_2 = 1.172, L_3 = 1.420, L_4 = 1.309.$$

Based on these limits, simultaneous 97.5% lower confidence limits are defined as follows. First,  $L_4$  and  $L_3$  are both greater than  $\lambda = 1.2$  and thus  $\tilde{L}_4 = \tilde{L}_3 = 1.2$ . However,  $L_2 < 1.2$  and thus  $\tilde{L}_2 = 1.172$  whereas the lower confidence bound for  $\lambda_1$  remains undefined.

### 3.3.2 Estimation of the maximum safe dose

Estimation methods introduced in Section 3.3.1 are easily extended to the problem of estimating the maximum safe dose (MAXSD) in dose-response studies with safety or toxicity endpoints. For consistency with 3.12 and 3.15 as well as notation used in the literature, it will be assumed that lower  $\mu_i$  is associated with higher toxicity or safety concerns. As in the previous section, we will focus on the ratio measure, i.e., define  $\lambda_i = \mu_i/\mu_0$ ,  $1 \leq i \leq m$ . Dose  $i$  is defined safe if  $\lambda_i > \lambda$  and unsafe otherwise. Here  $\lambda < 1$  and thus  $\lambda = 0.8$  indicates that a reduction of more than 20% in safety level (or, equivalently, increase in toxicity) is considered unacceptable.

The safety or toxicity response is generally expected to be monotone (as opposed to the efficacy response) and thus the following two definitions of MAXSD for any fixed  $\lambda < 1$  are equivalent:

$$\text{MAXSD} = \max\{i : \lambda_i > \lambda\}$$

and

$$\text{MAXSD} = \max\{i : \lambda_j > \lambda \text{ for all } 1 \leq j \leq i\}.$$

The hypotheses tested in this problem are the same as in 3.12. The only difference is that  $H_i$  now states that dose  $i$  is unsafe. Let

$$\widehat{\text{MAXSD}} = \max\{i : H_j \text{ is rejected for all } 1 \leq j \leq i\}$$

denote the estimated MAXSD. We are interested in controlling the FWER, i.e., in demonstrating that

$$P(\widehat{\text{MAXSD}} > \text{MAXSD}) \leq \alpha. \quad (3.17)$$

**TABLE 3.4:** Summary statistics for the daphnid length (mm) in the toxicology study.

	Treatment group					
	0	1	2	3	4	5
$n$	80	38	39	35	35	33
Mean	4.000	3.991	3.811	3.631	3.460	3.211
Pooled SD	0.174	0.174	0.174	0.174	0.174	0.174

To apply the SD1PC and SD2PC procedures to the problem of finding MAXSD, these procedures need to be modified as follows. Both procedures begin with  $H_1$  and proceed to  $H_2$  if  $H_1$  is rejected (i.e., dose 1 is declared safe), etc. However, if  $H_1$  cannot be rejected, the procedures automatically retain all hypotheses and all doses are declared unsafe. In this case, MAXSD is undefined. The SD1PC procedure rejects  $H_i$  if

$$t_{i,\max} = \max_{i \leq j \leq m} t_j > t_\alpha(k, \nu, \mathbf{R}_k),$$

where  $k = m - i + 1$  and  $\mathbf{R}_k = \{\rho_{ij}\}$ . Similarly, the SD2PC procedure rejects  $H_i$  if  $t_i > t_\alpha(\nu)$ . See Tamhane, Dunnett, Green and Wetherington (2001) for more information about the use of the SD1PC and SD2PC procedures in this problem.

**Estimation of MAXSD in a toxicology study**

To illustrate the use of the SD1PC and SD2PC procedures for estimating MAXSD, we consider a daphnid toxicology study (Tamhane et al., 2001). In this study, daphnids of the same age and genetic stock were randomly allocated to eight treatments: water control, solvent control and six concentrations of a pesticide (3.125, 6.25, 12.5, 25, 50 and 100 ppm). There were 40 daphnids in each group at the beginning of the study. Note that no significant difference between the two control groups was observed and the two groups were combined in the final analysis. Further, the 100 ppm dose group was excluded from the final analysis due to excessive mortality.

The endpoint of interest was response to toxicity measured in terms of growth, i.e., the lengths of the daphnids after 21 days of continuous exposure. Using a 15% reduction in the average length of a daphnid relative to placebo as a biologically significant threshold,  $\lambda$  was set to 0.85. Table 3.4 shows the summary statistics in the six remaining groups.

To apply the SD1PC and SD2PC procedures, we first need to compute the test statistics and critical values associated with a one-sided 0.05 level. The  $t$ -statistics computed using 3.15 are displayed in Table 3.5. Further, since the sample sizes in the dose groups were all approximately equal, let  $r_i = 80/36 = 2.22$  and  $\rho_{ij} = 0.85^2/(0.85^2 + 2.22) = 0.245$ . The critical values for the SD1PC

**TABLE 3.5:** Test statistics in the toxicology study.

	Comparison				
	1 vs. 0	2 vs. 0	3 vs. 0	4 vs. 0	5 vs. 0
$t$ -statistic	18.082	12.692	6.838	1.774	-5.505

procedure are given by

$$\begin{aligned} t_{0.05}(5, 254, 0.245) &= 2.307, \quad t_{0.05}(4, 254, 0.245) = 2.224, \\ t_{0.05}(3, 254, 0.245) &= 2.114, \quad t_{0.05}(2, 254, 0.245) = 1.952, \\ t_{0.05}(1, 254, 0.245) &= 1.652. \end{aligned}$$

The critical value for the SD2PC procedure is  $t_{0.05}(254) = 1.652$ .

As was stated above, the SD1PC procedure begins with  $H_1$ . It is easy to check that  $t_{1,\max} > 2.307$ ,  $t_{2,\max} > 2.224$  and  $t_{3,\max} > 2.114$  and thus the first three hypotheses in the sequence are rejected. However,  $t_{4,\max} < 1.952$ , which implies that  $H_4$  cannot be rejected. The last hypothesis in the sequence is automatically retained by the procedure and MAXSD is set to 3.

Considering the SD2PC procedure, the five hypotheses are again tested sequentially beginning with  $H_1$ . The procedure rejects the first four hypotheses in the sequence since the corresponding test statistics are greater than 1.652. The hypothesis  $H_5$  is retained since  $t_5 < 1.652$ , which leads to the conclusion that MAXSD = 4.

### 3.4 Power and sample size calculation for target dose estimation

Tamhane, Shi and Strassburger (2006) addressed the problem of designing an experiment to estimate MAXSD using the SD2PC procedure. These authors formulated the problem as follows. Consider the setup of Section 3.3.2. Define the power as the probability of rejecting all false  $H_i$ 's. The power requirement involves specification of two constants,  $\delta$  ( $0 < \delta < 1 - \lambda$ ) and  $1 - \beta$  ( $\alpha < 1 - \beta < 1$ ). These constants are incorporated in the following requirement:

$$P \left\{ \text{Reject all false } H_i \text{ with } \min_{j \leq i} \mu_j \geq (\lambda + \delta)\mu_0 \right\} \geq 1 - \beta. \quad (3.18)$$

For example, suppose  $\lambda = 0.85$ ,  $\delta = 0.05$  and  $1 - \beta = 0.80$ . Then any dose with mean  $\mu_i > 0.85\mu_0$  is safe, but we want to guarantee that all consecutive doses with means  $\mu_i > 0.90\mu_0$  are declared safe with probability at least 0.80.

Tamhane et al. (2006) showed that the power is minimized subject to

$$\mu_i \geq (\lambda + \delta)\mu_0 \text{ for } 1 \leq i \leq m, \quad \mu_{m+1} < (\lambda + \delta)\mu_0 \text{ for some } m \leq k$$

when  $m = k$ , i.e., all doses are safe and  $\mu_1 = \dots = \mu_k = (\lambda + \delta)\mu_0$ . Furthermore, when  $n_1 = \dots = n_k$  and  $n_0/n = r$ , this minimum power is given by

$$\int_0^\infty \left\{ \int_{-\infty}^\infty \Phi^k \left[ \frac{\tau z + c(u)}{\sqrt{1 - \tau^2}} \right] \phi(z) dz \right\} h_\nu(u) du, \quad (3.19)$$

where

$$-t_\alpha(\nu)u + \frac{\delta(\mu_0/\sigma)}{\sqrt{1/n + \lambda^2/n_0}}, \quad \tau = \frac{\lambda}{\sqrt{r + \lambda^2}}$$

and  $h_\nu(u)$  is the p.d.f. of  $U = \sqrt{\chi_\nu^2/\nu}$  and equals

$$h_\nu(u) = \frac{2(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} u^{\nu-1} e^{-\nu u^2/2}, \quad u \geq 0.$$

This expression can be used to compute the power for given  $n_0$ ,  $n$ ,  $k$ ,  $\delta$ ,  $\mu_0/\sigma$  and  $\alpha$ , or solve the dual problem of finding the optimum values of  $n_0$  and  $n$  for specified power,  $1 - \beta$ . This latter problem can be stated as the following optimization problem:

$$\text{Minimize } N = n_0 + kn \text{ subject to } 3.19 \geq 1 - \beta.$$

This problem can be solved using numerical search methods. Table 3.6 gives sample sizes  $n_0$  and  $n$  for selected values of other quantities.

These sample size calculations do not assume any knowledge of the shape of the dose-response function; i.e., the dose-response is a step function. If one assumes that the dose-response function is linear then significant savings are possible. The corresponding analytical expressions for power and a table of sample sizes are given in Tamhane et al. (2006). For example, for  $k = 3$ ,  $1 - \beta = 0.80$ ,  $\lambda = 0.90$  and  $\delta(\mu_0/\sigma) = 0.50$ , Table 3.6, which assumes that the dose-response is a step function, gives  $n_0 = 84$ ,  $n = 58$  and  $N = 258$ , while the linear response function assumption gives  $n_0 = 61$ ,  $n = 41$  and  $N = 184$  for a saving of 29%.

### 3.5 Hybrid approaches combining multiple testing and modeling

As mentioned in Section 3.1, an alternative approach to the multiple comparison procedures reviewed in Sections 3.2 and 3.3 is the use of model-based methods. Such methods assume a functional relationship between the response and the dose, taken as a quantitative factor, according to a pre-specified parametric model, such as a logistic,  $E_{\max}$  or log-linear model. The fitted model is then used to estimate an adequate dose to achieve a desired response. This

**TABLE 3.6:** Optimum sample sizes for finding MAXSD using the SD2PC procedure with  $\alpha = 0.05$  and  $\delta(\mu_0/\sigma) = 0.50$ .

$k$	$1 - \beta$	$\lambda$	$n_0$	$n$	$N$
3	0.70	0.80	58	47	199
		0.90	66	49	213
	0.80	0.80	72	56	240
		0.90	84	58	258
	0.90	0.80	93	70	303
		0.90	107	73	326
4	0.70	0.80	69	50	269
		0.90	79	52	287
	0.80	0.80	84	59	320
		0.90	97	61	341
	0.90	0.80	109	72	397
		0.90	125	75	425
5	0.70	0.80	76	53	341
		0.90	91	54	361
	0.80	0.80	95	61	400
		0.90	110	63	425
	0.90	0.80	121	74	491
		0.90	110	63	425

approach provides flexibility in interpolating the dose-response. However, the validity of these conclusions highly depends on the correct choice of the dose-response model, which is of course a priori unknown. This creates a dilemma in practice, because, within the regulated environment in which drug development takes place, it is required to have the analysis methods (including the choice of the dose-response model) defined prior to the study.

In the following we describe a hybrid strategy for the analysis of data from dose-response studies which combines multiple comparison procedures and modeling techniques. We review basic concepts on modeling techniques in Section 3.5.1. In Section 3.5.2 we discuss the integration of both strategies and describes the application of the so-called MCP-Mod approach to general linear models 3.6. In Section 3.5.3 we discuss a case study in great detail, illustrating relevant features related to the planning and analysis of clinical dose-finding studies.

### 3.5.1 Modeling techniques

In the context of modeling a continuous response  $y$  (which can be an efficacy or a safety variable) we assume a parallel group design with the dose levels  $d_1, d_2, \dots, d_m$  plus placebo  $d_0$ , for a total of  $m + 1 \geq 3$  arms. We start

by considering a one-way layout model

$$y_{ij} = f(d_i, \boldsymbol{\theta}) + \varepsilon_{ij}, \quad i = 0, 1, \dots, m, \quad j = 1, \dots, n_i, \tag{3.20}$$

where  $\boldsymbol{\theta}$  refers to the vector of model parameters,  $i$  to the dose group (as before,  $i = 0$  corresponds to placebo),  $j$  to the patient within dose group  $i$ , and  $\varepsilon_{ij}$ 's are independent normally distributed random errors with mean 0 and variance  $\sigma^2$ . Covariates and factorial treatment structures can be introduced as shown below in Section 3.5.2.

It often suffices to consider a standardized version of the dose-response model. That is, if the function  $f$  can be written as

$$f(d, \boldsymbol{\theta}) = \theta_0 + \theta_1 f^0(d, \boldsymbol{\theta}^0),$$

one needs to consider only  $f^0$ . For example, for the purpose of determining optimal contrasts to test for a significant dose-response effect later in Section 3.5.2, it is sufficient to consider the standardized model  $f^0$ . Table 3.7 lists a selection of models frequently used to describe dose-response relationships, together with their respective standardized versions. The corresponding shapes are illustrated in Figure 3.4.

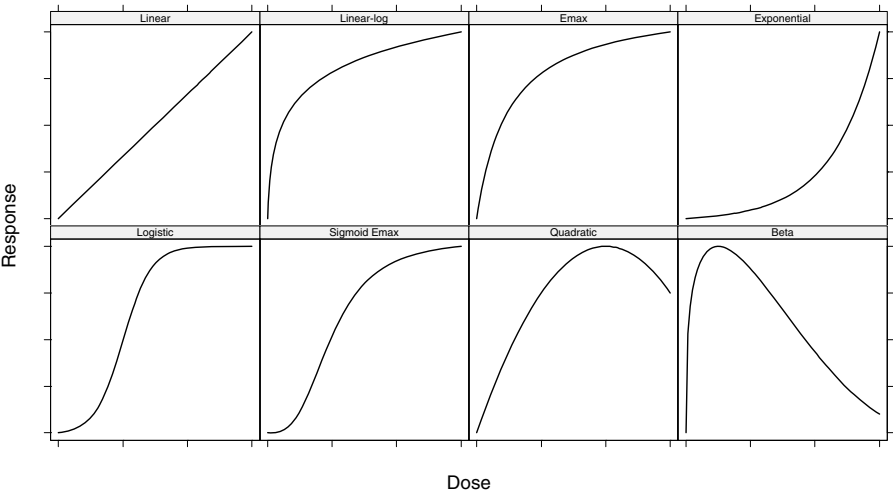


FIGURE 3.4: Typical dose-response shapes used in practice.

Since the specification of the model testing contrasts is the only information involving  $f$  needed prior to the start of the study, it suffices to consider strategies for producing best guesses for  $\boldsymbol{\theta}^0$ . Such estimates are typically derived from initial knowledge (or best guesses) of the expected percentage  $p^*$  of the maximum response associated with a given dose  $d^*$ . For example, in the  $E_{\max}$  model, an initial estimate for the single parameter in the standardized

**TABLE 3.7:** A selection of frequently used dose-response models.

Model	$f(d, \theta)$	$f^0(d, \theta^0)$
Linear	$E_0 + \delta d$	$d$
Linear log-dose	$E_0 + \delta \log(d + 1)$	$\log(d + 1)$
$E_{\max}$	$E_0 + E_{\max}d/(ED_{50} + d)$	$d/(ED_{50} + d)$
Exponential	$E_0 + E_1 \exp(d/\delta)$	$\exp(d/\delta)$
Logistic	$E_0 + E_{\max}/\{1 + \exp[(ED_{50} - d)/\delta]\}$	$1/\{1 + \exp[(ED_{50} - d)/\delta]\}$
Sigmoid $E_{\max}$	$E_0 + E_{\max}d^h/(ED_{50}^h + d^h)$	$d^h/(ED_{50}^h + d^h)$
Quadratic	$E_0 + \beta_1 d + \beta_2 d^2$	$d + (\beta_2/ \beta_1 )d^2$
Beta model	$E_0 + E_{\max}B(\alpha, \beta)(d/D)^\alpha(1 - d/D)^\beta$	$B(\alpha, \beta)(d/D)^\alpha(1 - d/D)^\beta$

model based on  $(d^*, p^*)$  is given by  $\widehat{ED}_{50} = d^*(1 - p^*)/p^*$ . Similar expressions for other models displayed Table 3.7 are given in Pinheiro, Bretz and Branson (2006). They described several linear and nonlinear regression dose response models commonly used in practice, including clinical interpretations for the associated model parameters.

Standard statistical software packages can be used to fit such non-linear regression models, which apply the Gauss-Newton algorithm for non-linear least squares, see the functions `nls` (Chambers and Hastie, 1992) and `gnls` (Pinheiro and Bates, 2000) in S-PLUS and R as well as PROC NLIN (Freund and Littell, 2000) in SAS. For a general account on non-linear regression we refer to Bates and Watts (1988).

### 3.5.2 MCP-Mod: An integrated dose-finding approach combining multiple comparisons and modeling

Recall from Section 3.2 that the power of standard hypotheses tests to detect a dose-response signal depends critically on the (unknown) dose-response relationship. Multiple contrast tests were proposed (Section 3.2.2) which lead to powerful and robust trend tests in general linear models. However, the optimal choice of the contrast coefficients remains an unsolved problem. In the Thalidomide example we have seen that contrast coefficients can often be related to specific dose-response shapes; see Table 3.1. Tukey et al. (1984) proposed to simultaneously use several trend tests based on different functional dose-response descriptions and to subsequently adjust the resulting  $p$ -values for multiplicity. Bretz et al. (2005) proposed an extension of this methodology, denoted MCP-Mod, which provides the flexibility of modeling for dose estimation, while preserving the robustness to model misspecification associated with multiple comparison procedures. Pinheiro, Bornkamp and Bretz (2006) discussed practical considerations regarding the implementation of this methodology. Extensions to Bayesian methods for estimating or selecting the dose-response curve from a sparse dose design have also been investigated (Neal, 2006; Wakana et al., 2007). Dette et al. (2008) constructed related optimal designs taking model uncertainty into account. A related MCPMod package has been implemented both in R (on CRAN at <http://www.r-project.org>) and

S-PLUS (on CSAN at <http://csan.insightful.com>); see Bornkamp et al. (2009) for details.

The MCP-Mod approach is implemented in five steps:

- Step 1.** Identify several candidate parametric models, which are likely to represent the underlying dose-response curve.
- Step 2.** Derive optimum contrast coefficients, such that the marginal power to detect a specific dose-response shape is maximized.
- Step 3.** Evaluate the significance of the individual models in terms of a multiple contrast test based on the previously derived optimal contrast coefficients.
- Step 4.** Select a candidate model associated with the most significant contrast test, or other model selection criteria such as AIC or BIC (provided significance has been shown in the previous step). Alternatively, multiple significant models can be selected if model averaging is preferred.
- Step 5.** Use the selected model(s) to produce inferences on adequate doses, employing a model-based approach.

In the following we describe the individual steps of the MCP-Mod approach in more detail by extending the results from Bretz et al. (2005) to the general linear model 3.6. To that end, the modeling framework assumed is that covariates enter the dose-response model linearly, that is

$$y_{ij} = f(d_i, \boldsymbol{\theta}, \mathbf{x}_{ij}) + \varepsilon_{ij} = \mathbf{x}_{ij}' \boldsymbol{\theta}_0 + \theta_1 f^0(d_i, \boldsymbol{\theta}^0), \quad (3.21)$$

where  $\mathbf{x}_{ij}$  denotes the covariate vector for patient  $j$  in dose group  $i$  and  $\boldsymbol{\theta}_0$  the associated vector of linear parameters. The remaining quantities are as defined before. The MCP-Mod approach will also be illustrated with a case study in Section 3.5.3.

## Step 1

Assume that we are given a set  $\mathcal{M} = \{M_\ell, \ell = 1, \dots, M\}$  of  $M$  candidate models. Each of these models is represented by a fixed mean vector  $\boldsymbol{\beta}_\ell^0 = (\beta_{\ell 0}^0, \beta_{\ell 1}^0, \dots, \beta_{\ell m}^0)'$ , which is the sub-vector of the parameter vector  $\boldsymbol{\beta}$  as derived from the standardized model  $f_\ell^0(d_i, \boldsymbol{\theta}_\ell^0) = \beta_{\ell i}^0$ ,  $i = 1, \dots, m$ . Note that  $\mathcal{M}$  may contain parameter specifications from different models  $f_\ell^0 \neq f_{\ell'}^0$  or include different parameter specifications  $\boldsymbol{\theta}_\ell^0 \neq \boldsymbol{\theta}_{\ell'}^0$  for the same model  $f_\ell^0 = f_{\ell'}^0$ ,  $1 \leq \ell \neq \ell' \leq M$ . In practice, the choice of the candidate model set  $\mathcal{M}$  in Step 1 requires a continuing interaction with the clinical team and may incorporate the input from pharmacometricians, clinicians, statisticians, etc.



## Step 2

Once the candidate model set  $\mathcal{M}$  has been identified, the goal is to select the best fitting model(s) while controlling the FWER. To this end, we test the null hypothesis  $H : \mathbf{c}'\boldsymbol{\beta}^0 = 0$  against the one-sided alternative  $K : \mathbf{c}'\boldsymbol{\beta}^0 > 0$  for a given  $(m+1) \times 1$  contrast vector  $\mathbf{c} = (c_0, c_1, \dots, c_m)'$  of known constants subject to  $\sum_{i=0}^m c_i = 0$ . The modeling context considered in this and the next step is that of an ANCOVA model in which the doses are represented by indicator functions. Referring back to the model specification (3.21), the regression matrix  $\mathbf{X}$  in the ANCOVA model would have rows given by  $\mathbf{x}_{ij}$  concatenated with a vector of length  $m$  with 1 in the  $i$ th position and 0 in the remaining (indicating that the patient is in dose group  $i$ ).

This leads to the construction of a single contrast test

$$t = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}}^0}{\hat{\sigma}\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})_0^{-}\mathbf{c}}},$$

where  $(\mathbf{X}'\mathbf{X})_0^{-}$  denotes the sub-matrix of  $(\mathbf{X}'\mathbf{X})^{-}$  associated with  $\boldsymbol{\beta}^0$ . Under the assumptions of model 3.6 and  $H$ , the test statistic  $t$  follows a central  $t$  distribution with  $\nu$  degrees of freedom. If  $H$  is not true,  $t$  follows a non-central  $t$  distribution with non-centrality parameter

$$\tau = \frac{\mathbf{c}'\boldsymbol{\beta}^0}{\sigma\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})_0^{-}\mathbf{c}}}.$$

Note that, at the planning stage of the trial, the parameters  $\boldsymbol{\beta}^0$  and  $\sigma$  are assumed known values. Thus, for a fixed design  $\mathbf{X}$ , the non-centrality parameter  $\tau$  depends only on the contrast vector  $\mathbf{c}$ . Consequently, to maximize the chance of rejecting  $H$  for a given model (i.e.,  $\boldsymbol{\beta}^0$ ), we select the contrast vector  $\mathbf{c}$  such that the non-centrality parameter  $\tau = \tau(\mathbf{c})$  is maximized. Using the Lagrange's multiplier device, one can show that for  $\boldsymbol{\Sigma} = (\mathbf{X}'\mathbf{X})_0^{-}$  the choice

$$\mathbf{c}_{opt} = \boldsymbol{\Sigma}^{-} \left( \boldsymbol{\beta}^0 - \frac{\boldsymbol{\beta}^{0'}\boldsymbol{\Sigma}^{-}\mathbf{1}}{\mathbf{1}'\boldsymbol{\Sigma}^{-}\mathbf{1}}\mathbf{1} \right) \quad (3.22)$$

maximizes  $\tau(\mathbf{c})$  and is thus optimal (Bornkamp, 2006). This solution is not unique since  $\tau(\mathbf{c}) = \tau(\gamma\mathbf{c})$  for all real  $\gamma > 0$ . Thus, if  $\tau(\mathbf{c})$  is maximized by  $\mathbf{c}_{opt}$ , so is  $\tau(\gamma\mathbf{c})$ . For uniqueness, we impose that the Euclidean length  $\|\mathbf{c}_{opt}\| = 1$ . Note that in the simplest case of the one-way ANOVA model 3.1, the standardized parameter vector  $\boldsymbol{\beta}^0$  reduces to the standardized mean vector  $\boldsymbol{\mu}^0 = (\mu_0^0, \dots, \mu_m^0)'$  and the covariance matrix  $\boldsymbol{\Sigma}^{-} = \text{diag}(n_0^{-1}, \dots, n_m^{-1})$  is diagonal. Thus, 3.22 simplifies to

$$\mathbf{c}_{opt} = \begin{pmatrix} n_0(\mu_0^0 - \bar{\mu}^0) \\ \vdots \\ n_m(\mu_m^0 - \bar{\mu}^0) \end{pmatrix},$$

where  $\bar{\mu}^0 = \sum_{i=0}^m n_i \mu_i^0 / \sum_{i=0}^m n_i$  is the overall standardized mean value (Bornkamp, 2006; Casella and Berger, 1990). In this case, the optimal contrast coefficients are simply computed by the group sample sizes  $n_i$  and the expected standardized mean responses  $\mu_i^0$ . In the balanced case the optimal contrast coefficients do not even depend on the sample sizes and the computation is further simplified.

### Step 3

Recall that the candidate set  $\mathcal{M}$  consists of  $M$  models. For each of the  $M$  models an optimal contrast vector  $\mathbf{c}_{opt,\ell}$  can then be computed from 3.22, which can be used to test for a significant dose-response effect. Our focus is thus on simultaneous inferences about the parameters  $\mathbf{c}'_{opt,\ell} \boldsymbol{\beta}^0$ . Let  $H_\ell : \mathbf{c}'_{opt,\ell} \boldsymbol{\beta}^0 = 0$  denote the resulting null hypothesis with the associated test statistic

$$t_\ell = \frac{\mathbf{c}'_{opt,\ell} \hat{\boldsymbol{\beta}}^0}{\hat{\sigma} \sqrt{\mathbf{c}'_{opt,\ell} (\mathbf{X}'\mathbf{X})_0^{-1} \mathbf{c}_{opt,\ell}}}, \quad \ell = 1, \dots, M.$$

One way to combine the test statistics  $t_\ell$  into a single decision rule is to consider the best contrast, that is, to consider  $t_{\max} = \max_\ell t_\ell$ , as discussed in 3.4. By construction of the individual hypotheses  $H_\ell$ , we therefore assess which parameter configuration best fits the available data. If  $t_{\max} > q$  for an appropriate critical value  $q$ , we conclude that the dose-response relationship is significant. If  $p_\ell$  denotes the associated multiplicity-adjusted  $p$ -value of  $t_\ell$ , an equivalent decision is obtained through  $\min_\ell p_\ell < \alpha$ . Otherwise, if  $t_{\max} < q$  (or  $\min_\ell p_\ell > \alpha$ ) no significant dose-related trend is detected and the MCP-Mod procedure stops without further conclusions.

Note that by performing several tests simultaneously the probability of committing at least one type I error increases with  $M$ . The computation of the critical value  $q$  should take this into account, such that the FWER is controlled at a pre-assigned level  $\alpha$ . We thus consider the joint distribution of the vector  $\mathbf{t}' = (t_1, \dots, t_M)$ . Under the assumptions above, the vector  $\mathbf{t}$  is  $M$ -variate  $t$  distributed with  $\nu$  degrees of freedom and correlation matrix  $\mathbf{R} = \mathbf{D}\mathbf{C}'(\mathbf{X}'\mathbf{X})_0^{-1}\mathbf{C}\mathbf{D}$ , where  $\mathbf{C}_{m \times M} = (\mathbf{c}_{opt,1}, \dots, \mathbf{c}_{opt,M})$  and  $\mathbf{D} = \text{diag}(\mathbf{c}'_{opt,\ell}(\mathbf{X}'\mathbf{X})_0^{-1}\mathbf{c}_{opt,\ell})_\ell^{-1/2}$ . The associated critical value is  $q = t_\alpha(M, \nu, \mathbf{R})$ , as introduced in Section 3.2.2. In the asymptotic case  $\nu \rightarrow \infty$  or if  $\sigma$  is known, the corresponding limiting multivariate normal distribution holds. Again, we refer to Genz and Bretz (2002, 2009) for a description of numerical integration methods required to compute the associated probabilities.

### Step 4

Once a significant trend has been shown, the MCP-Mod approach continues selecting a single model for the final target dose estimation. Recall from

the discussion at the end of Section 3.2.4 that maximum contrast tests are by construction consonant. Thus, if  $t_{\max}$  is significant, we know which individual contrast test is actually significant. The following approaches can be used:

- Select the model and parameter vector associated with the most significant test statistic.
- Construct a reference set  $\mathcal{M}^*$  of all significant models. Based on the reference set  $\mathcal{M}^*$ , one can then select a single model based on standard model selection criteria (AIC, BIC, etc) for the final target dose estimation step. Alternatively, one can keep  $\mathcal{M}^*$  and apply model averaging techniques (Buckland et al., 1997).

Note that the trend tests mentioned at the beginning of Section 3.2.2 do not provide any information beyond the assessment of whether a significant trend exists or not.

## Step 5

The final step consists of fitting the selected dose-response model(s) to the data and estimating the target dose(s) of interest using modeling techniques. For example, if the MINED is of interest, definitions 3.9 and 3.10 are extended to account for the continuous dose range  $(d_0, d_m]$ . The MINED associated with a model  $f(d, \theta)$  is defined as the smallest dose in  $(d_0, d_m]$  for which  $f(d, \theta) - f(d_0, \theta) \geq \delta$  (absolute scale) or  $f(d, \theta)/f(d_0, \theta) \geq \gamma$  (relative scale), where  $\delta > 0$  and  $\gamma > 1$  are clinically relevant effects and  $f(d_0, \theta)$  is assumed to be positive in the latter case. Those methods can be easily extended to safety endpoints and the estimation of the MAXSD. Further details on the estimation of target doses under MCP-Mod can be found in Bretz et al. (2005, 2008).

### 3.5.3 Asthma dose-finding trial

This section illustrates the use of MCP-Mod in the design and analysis of a dose-finding trial for a new compound for the indication of asthma. Four doses of the compound are investigated: 0.25, 0.5, 1, and 2 mg given orally once daily, plus a placebo and an active control. The primary clinical endpoint is the change from baseline in trough (24 hours post-dose) forced expiratory volume over 1 second ( $FEV_1$ ) after 14 days of treatment.

The main goals of the trial are twofold: to determine that the expected change in  $FEV_1$  varies with dose and, if so, to select a dose for the Phase III program (confirmatory trials) which produces a clinically relevant improvement over placebo of 120 mL in  $FEV_1$  change from baseline. MCP-Mod is utilized to address both of these goals.

To reduce the impact of inter-patient variability, which is known to be large for this indication, a crossover design was initially considered. However, the

**TABLE 3.8:** Candidate models in the asthma dose-finding trial.

Model	Representation
$E_{\max 1}$	$168.8d/(0.250 + d)$
$E_{\max 2}$	$153.9d/(0.053 + d)$
Linear	$75d$
Logistic	$-7.14 + 157.24/\{1 + \exp [(0.58 - d)/0.191]\}$
Quadratic	$232.92d - 90.2d^2$

relatively large number of treatments (6) and the required wash-out period (7 days) between treatments made a complete crossover unfeasible. The compromise was to use a balanced incomplete block crossover design in which each patient received 4 of the 6 treatments. The treatment sequences, in blocks of size 30, were determined prior to randomizing patients, so as to ensure balance in the order and combinations of treatments measured in the same patient.

Patients were accounted for as fixed effects in the dose-response model, together with treatment sequence, period and baseline FEV<sub>1</sub>. What is particularly interesting in this example is the fact that some of the covariates, such as patient and period, are known before the trial started and this information can therefore be utilized at the planning stage. We now consider each of the five MCP-Mod steps for this particular trial, thereby illustrating the general linear model formulation presented in Section 3.5.2

Data from the actual clinical trial is available for this example and is used to illustrate Steps 3, 4 and 5. The first two steps, which are implemented at the design stage and, therefore, before any data from the trial is available, do not use the observed data.

Step 1: Candidate models

After several discussions with the clinical team, five candidate models were selected to represent the anticipated dose-response shapes for the improvement in FEV<sub>1</sub> change from baseline over placebo:  $E_{\max}$  (2 models), linear, logistic, and quadratic. The parameter guesses for the candidate models were elicited through discussions with the clinical team and utilizing results from previous studies on the same compound and other drugs for the same indication. It was anticipated that the maximum treatment effect would be 150 ml improvement over placebo. The resulting candidate models are described in Table 3.8 and illustrated in Figure 3.5.

Step 2: Optimal model contrasts

Because of the presence of covariates, the more general formula 3.22 needs to be used to determine the optimal contrasts. That requires knowledge of the regression matrix **X** which depends on the number of patients to be recruited

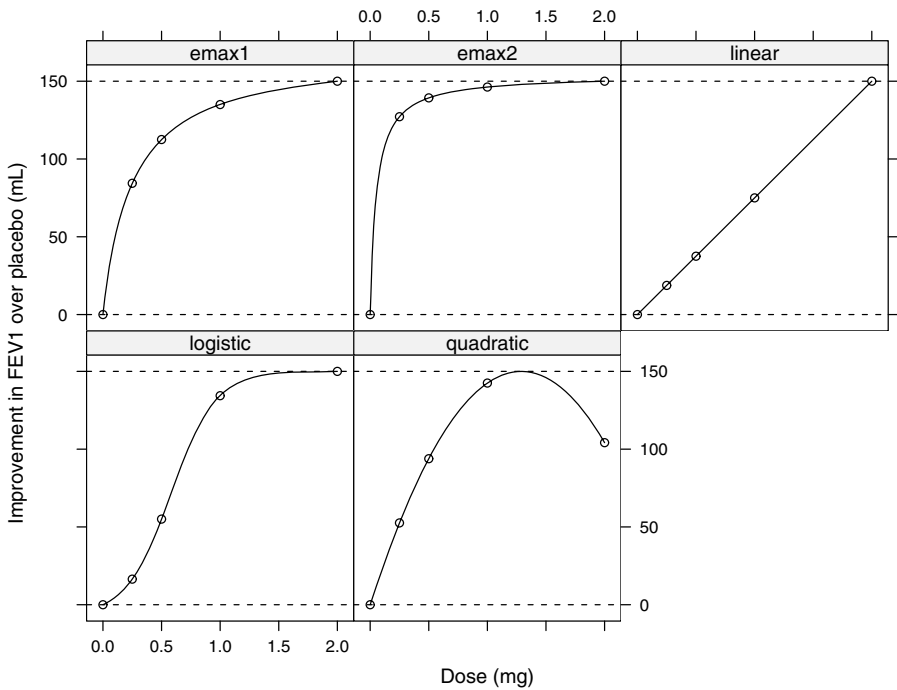


FIGURE 3.5: Candidate models in the asthma dose-finding trial.

for the trial. For this particular trial, the sample size was determined based on conservative pairwise tests with Bonferroni multiplicity correction, leading to  $N = 73$  patients. Based on the pre-defined treatment sequences, the corresponding  $\mathbf{X}$  matrix was derived and utilized with formula (3.22) to produce the optimal contrasts in Table 3.9. The  $\mathbf{X}$  matrix used in this step was obtained prior to the start of trial, depending only on the subset of covariates known at that stage, namely, patient, period, and dose. For illustration, the subset of  $\mathbf{X}$  corresponding to the first patient is shown below.

$$\begin{pmatrix} 1 & | & 1 & 0 & \cdots & 0 & | & 1 & 0 & 0 & 0 & | & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & | & 1 & 0 & \cdots & 0 & | & 0 & 1 & 0 & 0 & | & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & | & 1 & 0 & \cdots & 0 & | & 0 & 0 & 1 & 0 & | & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & | & 1 & 0 & \cdots & 0 & | & 0 & 0 & 0 & 1 & | & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The first column corresponds to the intercept, the next 73 columns to the patient indicators, followed by 4 columns of period indicators and 6 columns of dose indicators.

It is illustrative at this point to consider the correlation among the model contrasts, given in Table 3.10. All contrasts are highly positively correlated. Using the Bonferroni adjustment, one would need to use an adjusted signifi-

**TABLE 3.9:** Optimal model contrasts in the asthma dose-finding trial.

Dose	Model				
	$E_{\max 1}$	$E_{\max 2}$	Linear	Logistic	Quadratic
0	-0.815	-0.886	-0.477	-0.525	-0.724
0.25	-0.099	0.115	-0.316	-0.400	-0.236
0.5	0.141	0.217	-0.159	-0.119	0.142
1	0.329	0.267	0.162	0.471	0.589
2	0.445	0.288	0.789	0.573	0.228

**TABLE 3.10:** Correlations between optimal model contrasts in the asthma dose-finding trial.

	$E_{\max 2}$	Linear	Logistic	Quadratic
$E_{\max 1}$	0.957	0.803	0.861	0.929
$E_{\max 2}$		0.624	0.685	0.868
Linear			0.924	0.674
Logistic				0.866

cance level of  $\alpha/M$ , where  $M$  is the number of models, and thus the number of individual contrast tests ( $M = 5$  in our example). Accounting for the correlations between contrast tests by using the multivariate  $t$  distribution reduces the critical value substantially, as shown below.

Step 3: Testing for dose-response signal

From this step onward, the real data observed in the trial is used to illustrate the methods and derive results. Using the methods described in Section 3.5.2, the multiplicity-adjusted critical value at level  $\alpha = 0.05$  for the multiple contrasts test is  $t_\alpha(M, \nu, \mathbf{R}) = 2.019$ , which is considerably smaller than the Bonferroni-adjusted critical value of 2.344. The test statistics for the individual model contrasts, given in Table 3.11, were all highly significant, with corresponding  $p$ -values virtually equal to zero. As a result, the significance of the dose-response signal was established and all models were considered in the next step.

Step 4: Model selection

The AIC criterion was used to select the best model. Table 3.12 displays the AIC values for each fitted model, when available.

The logistic model did not converge for the observed data, so no AIC could be derived. Note also that, even though there are two  $E_{\max}$  models in the candidate set, only one  $E_{\max}$  fit can be obtained. Based on the AIC results, the  $E_{\max}$  model was chosen to represent the dose-response profile. The

**TABLE 3.11:** Test statistics for model contrast tests.

Model	$E_{\max 1}$	$E_{\max 2}$	Linear	Logistic	Quadratic
Test statistic	7.238	6.715	6.185	6.641	6.804

**TABLE 3.12:** AIC values for individual models.

Model	$E_{\max}$	Linear	Logistic	Quadratic
AIC	-472.63	-454.45	NA	-469.89

estimated improvement over placebo estimated from the fitted  $E_{\max}$  model was  $168d/(0.36 + d)$ .

### Step 5: Estimation of MINED

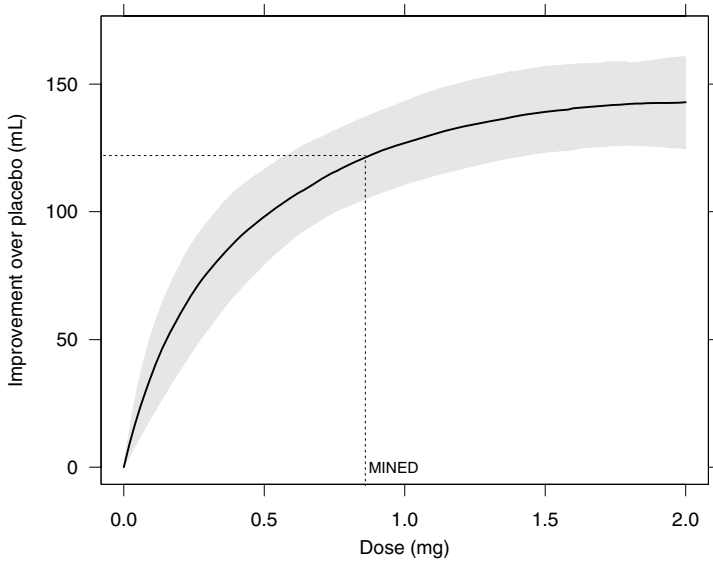
Based on the fitted  $E_{\max}$  of Step 4, the smallest dose giving the clinically relevant improvement over placebo of 120 mL is estimated to be 0.84 mg. This is the MINED estimate produced by MCP-Mod in this trial. The precision of the MINED estimate was evaluated via a non-parametric bootstrap approach: patients were sampled with replacement within the same treatment sequence group (to preserve the same design as in the original data) and the full MCP-Mod procedure was applied to the resampled data, producing a bootstrap sample of MINED values. The 90% confidence interval for the MINED, corresponding to the 5% and 95% quantiles of the bootstrap sample, was  $[0.48, 1.45]$ , reflecting the uncertainty about the estimate.

The bootstrap samples were also used to evaluate the precision of the estimated dose response: predicted values were obtained for each sample, and pointwise 90% confidence intervals based on quantiles were obtained. [Figure 3.6](#) shows the median predicted values for the improvement over placebo, with corresponding 90% pointwise confidence bands.

The use of MCP-Mod in this trial provided answers to both key questions that the clinical team wanted to address in it: establishing dose-response signal and selecting a dose for the confirmatory phase. The former was unequivocally shown by the large test statistics for the model contrasts, while the latter was provided by modeling the dose-response profile. The ultimate success of the dose selection will need to wait until the Phase III program is concluded.

### Acknowledgements

Ajit C. Tamhane's research was supported by grants from the National Heart, Lung and Blood Institute.



**FIGURE 3.6:** Estimated dose-response model with bootstrap 90% pointwise confidence bands. MINED corresponding to the clinically relevant improvement of 120 mL over placebo indicated in the plot.



# Chapter 4

---

## *Analysis of Multiple Endpoints in Clinical Trials*

Ajit C. Tamhane

*Northwestern University*

Alex Dmitrienko

*Eli Lilly and Company*

---

### 4.1 Introduction

Most human diseases are characterized by multidimensional etiology and the efficacy of an experimental treatment frequently needs to be assessed on multiple outcome measures — commonly referred to as endpoints. There is a variety of ways in which the contribution of each endpoint can be accounted for in the primary analysis, for example, the trial's sponsor can treat endpoints as independent entities or as manifestations of a single underlying cause. The following examples illustrate the common approaches to analyzing multiple endpoints.

**Example 1.** If each endpoint independently provides a proof of efficacy, the trial's outcome is declared positive if at least one endpoint is associated with a significant improvement compared to the control. Gong, Pinheiro and DeMets (2000) gave several examples of cardiovascular trials in which the primary objective had two components (primary endpoint and principal secondary endpoint), e.g., mortality and mortality plus morbidity due to heart failure in the VEST trial (Cohn et al., 1998) and mortality plus cardiovascular morbidity and mortality in the PRAISE-I trial (Packer et al., 1996).

**Example 2.** Another commonly used approach to defining the primary objective in clinical trials with multiple endpoints is based on the development of *composite endpoints*. A composite endpoint can be based on a sum of multiple scores or combination of multiple events. In the case of multiple events, a patient achieves a composite endpoint if he or she experiences any of the pre-specified events related to morbidity or mor-

tality. For example, the primary objective of the Losartan Intervention For Endpoint reduction (LIFE) trial was to study the effect of losartan on the composite endpoint of cardiovascular death, myocardial infarction, and stroke (Dahlöf et al., 2002).

**Example 3.** When the multiple endpoints are biologically related to each other (for example, they measure different aspects of the same underlying cause), the primary objective can be defined in terms of a combination of individual effects across the endpoints. The mitoxantrone trial in patients with progressive multiple sclerosis evaluated the overall effect of five clinical measures: expanded disability status scale, ambulation index, number of treated relapses, time to first treated relapse and standardized neurological status (Hartung et al., 2002). The analysis was performed based on a multivariate approach that accounted for the correlations among the endpoints.

**Example 4.** In certain areas, a clinically meaningful effect is defined as the simultaneous improvement in multiple measures. In this case, the primary objective of a clinical trial is met if the test drug shows a significant effect with respect to all the endpoints. Offen et al. (2007) gave a list of more than 15 examples in which a positive trial is defined in terms of two or more significant endpoints. They included clinical trials for the treatment of migraine, Alzheimer's disease and osteoarthritis.

Due to an increasing number of studies dealing with conditions that exhibit a complex etiology, there has been much attention in the clinical trial literature on the development of statistical methods for analyzing multiple endpoints. O'Brien's (1984) paper ignited a flurry of research in this area and a number of procedures for testing multiple endpoints have been developed over the past two decades. O'Neill (2006) mentions multiple endpoints as one of the key areas for guidance development in FDA's critical path initiative and writes that "The many statistical methods and approaches that have been suggested in the literature in recent years now deserve to be digested and placed in the context of how they can be best used in clinical trials." The purpose of this chapter is to attempt to answer this need by giving a critical overview of the various procedures along with their assumptions, pros and cons, and domains of applicability. For some other reviews of the literature on multiple endpoints, the reader is referred to Chi (1998), Comelli and Klersy (1996), Geller (2004), Huque and Sankoh (1997), Kieser, Reitmeir and Wassmer (1995), Sankoh, Huque and Dubey (1997) and Zhang, Quan, Ng and Stepanavage (1997).

This chapter is organized as follows. Section 4.2 introduces inferential goals and main classes of multiple endpoint procedures. It is assumed in this chapter that all endpoints are primary or co-primary. Gatekeeping procedures that deal with endpoints classified into hierarchically ordered categories such as primary and secondary will be discussed in Chapter 5. We begin with an overview of multiple testing procedures for the assessment of the treatment's

effect on each individual endpoint (at-least-one procedures) in Section 4.3. Global procedures aimed at examining the overall efficacy of the treatment are reviewed in Section 4.4. Section 4.5 discusses all-or-none procedures arising in clinical trials when the treatment's efficacy needs to be demonstrated for all endpoints. Section 4.6 describes the superiority-noninferiority approach which requires demonstrating the treatment's superiority on at least one endpoint and noninferiority on others. Finally, Section 4.7 discusses software implementation of selected procedures for testing multiple endpoints in clinical trials.

## 4.2 Inferential goals

To define inferential goals of multiple endpoint procedures, we will consider a clinical trial with two treatment groups. The trial's objective is to assess the effect of the experimental treatment (hereafter referred to simply as the treatment) on  $m$  endpoints compared to that of the control, e.g., placebo. Let  $\delta_i$  denote an appropriate measure of the true treatment effect for the  $i$ th endpoint (e.g., mean difference, odds ratio or log hazard ratio). The efficacy of the treatment on the  $i$ th endpoint is assessed by testing the hypothesis of no treatment difference. The multiple testing setting gives rise to the problem of Type I error rate inflation. Depending on the trial's objectives, this problem can be addressed by performing a multiplicity adjustment, combining evidence of the treatment's effect across the endpoints (e.g., by combining multiple endpoints into a single composite endpoint) or utilizing other approaches. This section gives a brief overview of four main approaches to the analysis of multiple endpoints.

### 4.2.1 At-least-one procedures

If each multiple endpoint is independently clinically relevant (and can potentially be associated with its own regulatory claim), the multiple endpoint problem can be formulated as a multiple testing problem. Cardiovascular clinical trials listed in Example 1 in the Introduction serve as an illustration of this approach. In these trials the endpoints are treated as clinically independent entities and the sponsor is interested in assessing the effect of the experimental treatment on each endpoint.

Given this structure of the primary objective, the trial is declared positive if at least one significant effect is detected. The global hypothesis is defined as the intersection of hypotheses for individual endpoints and the testing problem is stated as

$$H_I = \bigcap_{i=1}^m (\delta_i \leq 0) \text{ versus } K_U = \bigcup_{i=1}^m (\delta_i > 0). \quad (4.1)$$

The global hypothesis is rejected if one or more individual hypotheses of no treatment effect are demonstrated to be false. This is a prototypical multiple testing problem, known as the *union-intersection* (UI) problem (see Section 2.3.1), which requires a *multiplicity adjustment*. The objective of a multiplicity adjustment is to control the familywise error rate (FWER),

$$\text{FWER} = P\{\text{Reject at least one true hypothesis}\},$$

at a designated level  $\alpha$  by adjusting the level of each test downward. As in other clinical trial applications, FWER needs to be controlled in the strong sense; i.e., it must not be greater than  $\alpha$  regardless of which hypotheses are true and which are false (Hochberg and Tamhane, 1987). Procedures that can be used in this multiple endpoint problem (at-least-one procedures) are described in Section 4.3.

It is worth noting that the false discovery rate (FDR) proposed by Benjamini and Hochberg (1995) is generally not appropriate for the multiple endpoint problem for the following reasons.

- FDR is suitable for testing a large number of hypotheses whereas the number of endpoints is generally very small.
- FDR is suitable for exploratory studies in which a less stringent requirement of control of the proportion of false positives is acceptable. However, tests for multiple endpoints are generally confirmatory in nature required for drug approval and labeling.
- The problem of testing multiple endpoints often has additional complications such as ordered categories of endpoints, e.g., primary, co-primary and secondary with logical restrictions, and decision rules based on tests for superiority and noninferiority on different endpoints. The FDR approach is not designed to handle such complex decision rules.

### 4.2.2 Global procedures

In many clinical trial applications it is desired to show that the treatment has an *overall* effect across the endpoints without necessarily a large significant effect on any one endpoint (see Examples 2 and 3 in the Introduction).

To establish an overall treatment effect, usually a point null hypothesis of no difference between the treatment and control is tested against a one-sided alternative:

$$H_I^* : \delta_i = 0 \text{ for all } i \text{ versus } K_U^* : \delta_i \geq 0 \text{ for all } i \text{ and } \delta_i > 0 \text{ for some } i. \quad (4.2)$$

It is well-known (O'Brien, 1984) that Hotelling's  $T^2$ -test is inappropriate for this problem because it is a two-sided test, and hence lacks power for detecting one-sided alternatives necessary for showing the treatment efficacy.

In this case one-sided *global procedures* that have been proposed as alternatives to the  $T^2$ -test are more appropriate.

Global procedures are conceptually similar to *composite endpoints* (defined in Example 2) in that they also rely on reducing the number of dimensions in a multiple endpoint problem by combining multiple measures into a single one, e.g., combining tests for individual endpoints into a single test. However, unlike composite endpoints based on a sum or other simple function of individual scores, global testing procedures address the interdependence of multiple endpoints, i.e., they account for the correlations among them. For a detailed discussion of global procedures, see [Section 4.4](#).

### 4.2.3 All-or-none procedures

Another formulation of the multiple endpoint problem pertains to the requirement that the treatment be effective on *all* endpoints (see Example 4 in the Introduction). This problem is referred to as the *reverse multiplicity problem* by Offen et al. (2007) and represents the most stringent inferential goal for multiple endpoints. In mathematical terms, this is an example of an *intersection-union* (IU) problem introduced in Section 2.3.2. Within the IU framework, the global hypothesis is defined as the union of hypotheses corresponding to individual endpoints and thus the testing problem is stated as

$$H_U = \bigcup_{i=1}^m (\delta_i \leq 0) \text{ versus } K_I = \bigcap_{i=1}^m (\delta_i > 0). \quad (4.3)$$

To reject  $H_U$ , one needs to show that all individual hypotheses are false (the treatment effects for all endpoints are significant). All-or-none testing procedures are discussed in Section 4.5.

### 4.2.4 Superiority-noninferiority procedures

The hybrid superiority-noninferiority testing approach to testing multiple endpoints provides a viable alternative to the stringent all-or-none testing approach described in Section 4.2.3. In this case, the goal is to demonstrate that the treatment is superior to the control on at least one endpoint and not inferior on all other endpoints.

To define the null and alternative hypotheses for this setting, let  $\eta_k \geq 0$  denote the superiority threshold (commonly,  $\eta_k = 0$ ) and  $\varepsilon_k > 0$  denote the noninferiority threshold for the  $k$ th endpoint. In other words, the treatment is superior to the control on the  $k$ th endpoint if  $\delta_k > \eta_k$  and noninferior if  $\delta_k > -\varepsilon_k$ ,  $k = 1, \dots, m$ .

For the  $k$ th endpoint, the superiority testing problem is stated as

$$H_k^{(S)} : \delta_k \leq \eta_k \text{ versus } K_k^{(S)} : \delta_k > \eta_k.$$

Similarly, the noninferiority testing problem for the  $k$ th endpoint is stated as

$$H_k^{(N)} : \delta_k \leq -\varepsilon_k \text{ versus } K_k^{(N)} : \delta_k > -\varepsilon_k.$$

Note that the difference between  $H_k^{(S)}$  and  $H_k^{(N)}$  is simply a shift. Showing superiority involves clearing a higher bar than showing noninferiority.

The overall superiority testing problem is given by

$$H_I^{(S)} = \bigcap_{k=1}^m H_k^{(S)} \text{ versus } K_U^{(S)} = \bigcup_{k=1}^m K_k^{(S)}.$$

The global superiority hypothesis  $H_I^{(S)}$  is rejected if *at least one*  $H_k^{(S)}$  is rejected. Similarly, the overall noninferiority testing problem is given by

$$H_U^{(N)} = \bigcup_{k=1}^m H_k^{(N)} \text{ versus } K_I^{(N)} = \bigcap_{k=1}^m K_k^{(N)}.$$

The global noninferiority hypothesis  $H_U^{(N)}$  is rejected if *all*  $H_k^{(N)}$  are rejected. To combine the two global hypotheses and formulate the superiority-noninferiority testing approach, we need to consider testing the union of the global superiority and global noninferiority hypotheses

$$H_U^{(SN)} = H_I^{(S)} \cup H_U^{(N)} \text{ versus } K_I^{(SN)} = K_U^{(S)} \cap K_I^{(N)}.$$

In other words, the trial's objective is met if there is superior efficacy for at least one endpoint ( $K_U^{(S)}$ ) and noninferior efficacy for all endpoints ( $K_I^{(N)}$ ). This superiority-noninferiority testing problem becomes equivalent to the all-or-none testing problem (Section 4.2.3) if the noninferiority margins are set to 0 for all endpoints, i.e.,  $\varepsilon_1 = \dots = \varepsilon_m = 0$ . Superiority-noninferiority procedures are discussed in Section 4.6.

### 4.3 At-least-one procedures

This section describes  $p$ -value-based procedures as well as parametric (normal theory) and resampling-based procedures for multiple endpoints when it is desired to demonstrate the treatment's superiority on at least one endpoint.

#### 4.3.1 Procedures based on univariate $p$ -values

Frequently, different test statistics (e.g.,  $t$ -statistics, log-rank statistics, chi-square statistics, etc.) are used to compare the treatment and control groups on different endpoints because of the different scales on which the endpoints

are measured. A standard approach to put the results of these different tests on the same scale is via their  $p$ -values. Therefore  $p$ -value-based procedures are of interest in the analysis of multiple endpoints. Section 2.6 provides a detailed description of popular multiple testing procedures based on univariate  $p$ -values. Here we briefly review basic procedures from that section with emphasis on issues specific to multiple endpoints. In addition, we describe two  $p$ -value-based procedures introduced specifically for the multiple endpoint problem.

## Bonferroni and related procedures

The simplest of the  $p$ -value-based procedures is the well-known Bonferroni procedure introduced in Section 2.6.1. This procedure is known to be conservative especially when there are many endpoints and/or they are highly correlated. More powerful stepwise versions of the Bonferroni procedure, e.g., the Holm and Hochberg procedures, are also described in Section 2.6.

The basic Bonferroni procedure allocates the Type I error rate  $\alpha$  equally among the endpoints. A weighted version of this procedure allows unequal allocation which is useful for unequally important endpoints, e.g., a clinical trial with a principal secondary endpoint which may provide the basis for a new regulatory claim. The weighted Bonferroni procedure can also be employed in trials where some endpoints are adequately powered and the others are underpowered.

As an example, consider the Carvedilol cardiovascular trials (Fisher and Moyé, 1999) in which exercise capability plus quality of life served as the primary endpoint and mortality was a key secondary endpoint. The primary endpoint was not significant at the 0.05 level in the trials with the exercise capability endpoint while the secondary mortality endpoint was highly significant in the combined analysis across the trials. Although the problem of interpreting the results of such trials is a vexing one (see O'Neill (1997) and Davis (1997) for contrasting views), we will assume, for the sake of illustration, that the mortality endpoint was prospectively defined as a co-primary endpoint.

In this case a decision rule based on the weighted Bonferroni procedure can be set up. In the general case of  $m$  endpoints, the rule uses an additive alpha allocation scheme. Let  $w_1, \dots, w_m$  be positive weights representing the importance of the endpoints such that they sum to 1. The hypothesis of no treatment effect for the  $i$ th endpoint is tested at level  $\alpha_i$ , where  $\alpha_i = w_i\alpha$  and thus

$$\sum_{i=1}^m \alpha_i = \alpha.$$

In the Carvedilol example, a Bonferroni-based decision rule could have been constructed by assigning a large fraction of the Type I error rate  $\alpha$  to the

exercise capability endpoint (i.e., by choosing  $\alpha_1$  close to  $\alpha$ ) and “spending” the rest on the mortality endpoint ( $\alpha_2 = \alpha - \alpha_1$ ).

A slightly sharpened version of this rule, termed the *prospective alpha allocation scheme* (PAAS) method, was proposed by Moyé (2000). Assuming that the  $p$ -values for the individual endpoints are independent, we have

$$\prod_{i=1}^m (1 - \alpha_i) = 1 - \alpha.$$

Moyé’s solution to the problem of two co-primary endpoints was to select the fraction of the Type I error rate  $\alpha$  allocated to the primary endpoint and calculate the significance level for the other endpoint from the above identity. Specifically, let  $0 < \alpha_1 < \alpha$  and

$$\alpha_2 = 1 - \frac{1 - \alpha}{1 - \alpha_1}.$$

In the context of the Carvedilol example, if  $\alpha = 0.05$  and  $\alpha_1 = 0.045$ , then  $\alpha_2 = 0.0052$ , which is only marginally larger than  $\alpha_2 = 0.005$  given by the weighted Bonferroni allocation.

## Fixed-sequence procedure

Maurer, Hothorn and Lehmacher (1995) considered clinical trials in which the endpoints are *a priori* ordered (e.g., in terms of their importance). They applied a fixed-sequence method that tests the ordered endpoints sequentially at level  $\alpha$  as long as the test is significant and stops testing when a non-significant result is encountered. Effectively,  $\alpha$  is unused when the procedure rejects the hypothesis of no treatment effect for an endpoint and thus it can be carried over to the next endpoint. All  $\alpha$  is used up when no treatment effect is detected with none left for the remaining endpoints. We refer to this as the “use it or lose it” principle.

The fixed-sequence testing approach is widely used in clinical trials and endorsed by regulatory agencies; see, for example, the CPMP guidance document on multiplicity issues (CPMP, 2002). As an example, this testing approach was adopted in the adalimumab trial in patients with rheumatoid arthritis (Keystone et al., 2004). The trial included three endpoints (American College of Rheumatology response rate, modified total Sharp score and Health Assessment Questionnaire score) that were prospectively ordered and tested sequentially. Since each test was carried out at an unadjusted  $\alpha$  level, this approach helped the trial’s sponsor maximize the power of each individual test. Note, however, that the overall power of the fixed-sequence procedure depends heavily on the true effect size of the earlier endpoints. The power of the procedure is increased if the likelihood of detecting a treatment effect for the endpoints at the beginning of the sequence is high. On the other hand, if at least one of the earlier endpoints is underpowered, the procedure is likely



to stop early and miss an opportunity to evaluate potentially useful endpoints later in the sequence. To improve the power of the fixed-sequence procedure, it is critical to order the endpoints based on the expected strength of evidence beginning with the endpoints associated with the largest effect size (Huque and Alosch, 2008).

## Fallback procedure

A useful generalization of the fixed-sequence procedure was proposed by Wiens (2003). Wiens constructed the fallback procedure by allocating pre-specified fractions,  $w_1, \dots, w_m$ , of  $\alpha$  to the  $m$  *a priori* ordered endpoints subject to

$$\sum_{i=1}^m w_i = 1.$$

The procedure begins with the first endpoint in the sequence which is tested at level  $\alpha_1 = \alpha w_1$ . Further, the  $i$ th endpoint is tested at level  $\alpha_i = \alpha_{i-1} + \alpha w_i$  if the previous endpoint is significant and level  $\alpha_i = \alpha w_i$  otherwise. In other words, if a certain test is not significant, its significance level ( $\alpha_i$ ) is used up and, if it is significant, its level is carried over to the next endpoint, hence the name *fallback procedure*. Note that this procedure also uses the “use it or lose it” principle.

The fallback procedure is uniformly more powerful than the Bonferroni procedure and reduces to the fixed-sequence procedure if all Type I error rate is spent on the first endpoint in the sequence, i.e.,  $w_1 = 1$  and  $w_2 = \dots = w_m = 0$ . The advantage of the fallback procedure is that one can continue testing even when the current test is not significant in contrast to the fixed-sequence procedure which stops testing as soon as it encounters a nonsignificant result.

As an illustration, consider a clinical trial with two endpoints, the first of which (functional capacity endpoint) is adequately powered and the other one (mortality endpoint) is not (Wiens, 2003). Wiens computed the power of the fallback procedure in this example assuming that  $w_1 = 0.8$  and  $w_2 = 0.2$  (80% of the Type I error rate is spent on the functional capacity endpoint and 20% on the mortality endpoint) and the two-sided  $\alpha = 0.05$ . Under this weight allocation scheme, the power for the mortality endpoint was substantially improved (from 50% to 88% compared to the Bonferroni procedure with the same set of weights) whereas the power for the functional capacity endpoint was reduced by a trivial amount (from 95% to 94%).

The overall power of the fallback procedure is heavily influenced by the effect sizes of the ordered endpoints and the significance levels for their tests (or, equivalently, the pre-specified weights). As shown in the next section, the power can be improved by arranging the endpoints in terms of the expected effect size, i.e., from the largest effect size to the smallest effect size. In addition, the expected effect size can help determine the significance levels. For

example, Huque and Alosch (2008) recommended defining the significance levels proportional to the reciprocals of the effect sizes. This choice helps increase the power of the early tests which will, in turn, raise the significance levels for the endpoints toward the end of the sequence.

## Comparison of the fixed-sequence and fallback procedures

To assess the robustness of the fixed-sequence and fallback procedures with respect to the monotonicity assumption, a simulation study was conducted. A clinical trial with two arms (treatment and placebo) was simulated. The treatment-placebo comparison was performed for three ordered endpoints (Endpoints 1, 2 and 3). The endpoints were tested sequentially, beginning with Endpoint 1, using the fixed-sequence method at the one-sided 0.025 level. The endpoint outcomes were assumed to follow a multivariate normal distribution with a compound-symmetric correlation matrix (i.e., the outcomes were equicorrelated). The sample size per group ( $n = 98$ ) was chosen to achieve 80% power for each univariate test when the true effect size is 0.4. The calculations were performed using 10,000 replications.

The power of the fixed-sequence and fallback procedures for the three endpoint tests is displayed in Table 4.1 for three values of the common correlation coefficient ( $\rho = 0, 0.2$  and  $0.5$ ) and three sets of endpoint-specific effect sizes,  $e_i$  ( $i = 1, 2, 3$ ). The following three scenarios were considered:

- Scenario 1. All tests are adequately powered,  $e_1 = 0.4$ ,  $e_2 = 0.4$ ,  $e_3 = 0.4$ .
- Scenario 2. The first test is underpowered but the other tests are adequately powered,  $e_1 = 0.3$ ,  $e_2 = 0.4$ ,  $e_3 = 0.4$ .
- Scenario 3. The first test is overpowered but the other tests are adequately powered,  $e_1 = 0.5$ ,  $e_2 = 0.4$ ,  $e_3 = 0.4$ .

Consider first the case of a constant effect size (Scenario 1 in Table 4.1). Since each test serves as a gatekeeper for the tests placed later in the sequence, the power of the individual tests in the fixed-sequence procedure declines fairly quickly as one progresses through the sequence. While the power of the first test is equal to its nominal value (80%), the power of the last test drops to 61% when the endpoints are moderately correlated ( $\rho = 0.5$ ). A greater power loss is observed with the decreasing correlation among the endpoints. Furthermore, the fixed-sequence procedure is quite sensitive to the assumption that the true ordering of the endpoints (in terms of the effect sizes) is close to the actual ordering. If the first test is underpowered (Scenario 2), it creates a “domino effect” that suppresses the power of the other tests. Comparing Scenario 2 to Scenario 1, the power of the last test decreases from 61% to 46% for moderately correlated endpoints and from 51% to 35% for uncorrelated endpoints. In general, the power of the fixed-sequence procedure is maximized if the outcome of the first test is very likely to be significant (see Westfall and Krishen, 2001; Huque and Alosch, 2008). This property of the fixed-sequence

**TABLE 4.1:** Power of the fixed-sequence and fallback procedures in a clinical trial with three endpoints as a function of the effect sizes and correlation. The fixed-sequence and fallback procedures are carried out at the one-sided 0.025 level. The weighting scheme for the fallback procedure is  $w_1 = 0.5$ ,  $w_2 = 0.25$  and  $w_3 = 0.25$ .

Correlation	Power of individual tests (%) (Endpoint 1, Endpoint 2, Endpoint 3)	
	Fixed-sequence procedure	Fallback procedure
Scenario 1 ( $e_1 = 0.4, e_2 = 0.4, e_3 = 0.4$ )		
0	(79.6, 63.4, 50.8)	(69.5, 72.3, 73.4)
0.2	(79.6, 65.2, 54.4)	(69.5, 71.7, 72.5)
0.5	(79.6, 68.1, 61.0)	(69.5, 70.7, 72.0)
Scenario 2 ( $e_1 = 0.3, e_2 = 0.4, e_3 = 0.4$ )		
0	(54.9, 43.9, 35.2)	(43.2, 68.3, 71.3)
0.2	(54.9, 46.1, 39.1)	(43.2, 67.5, 70.4)
0.5	(54.9, 49.4, 45.7)	(43.2, 66.1, 69.5)
Scenario 3 ( $e_1 = 0.5, e_2 = 0.4, e_3 = 0.4$ )		
0	(94.0, 75.0, 60.0)	(89.9, 75.0, 75.0)
0.2	(94.0, 75.7, 62.4)	(89.9, 74.8, 74.2)
0.5	(94.0, 77.1, 67.2)	(89.9, 74.7, 74.2)

procedure is illustrated in Scenario 3. It takes an overpowered test at the beginning of the sequence to bring the power of the other tests closer to its nominal level. For example, the power of the second test in the fixed-sequence procedure is only three to five percentage points lower than the nominal value (75-77% versus 80%) when the procedure successfully passes the first test 94% of the time.

Further, consider the properties of the fallback procedure based on the following weighting scheme for the three tests:  $w_1 = 0.5$ ,  $w_2 = 0.25$  and  $w_3 = 0.25$ . The power of the first test in the fallback procedure is uniformly lower across the three scenarios compared to the fixed-sequence procedure. This is due to the fact that the fallback procedure, unlike the fixed-sequence procedure, spends only half of the available Type I error rate on the first endpoint. The remaining fraction is distributed over the other two tests, which leads to a substantial improvement in their power in Scenarios 1 and 2. Specifically, the power of the second and third tests for the fallback procedure is much closer to the nominal level (80%) compared to the fixed-sequence procedure. Note also that, in the case of the fallback procedure, the power of individual tests stays at a constant level or increases toward the end of the sequence in all three scenarios (even when the monotonicity assumption is violated). Finally, while the power of tests placed later in the sequence improves with the increasing correlation for the fixed-sequence procedure, the fallback procedure exhibits an opposite trend. The power of the second and third tests declines slowly as the correlation among the endpoints increases

(the difference becomes very small when the first test is overpowered as in Scenario 3).

## Adaptive alpha allocation approach

Li and Mehrotra (2008) proposed a multiple testing procedure, which they referred to as the *adaptive alpha allocation approach* or *4A procedure*. Consider a clinical trial with  $m$  endpoints and assume that the endpoints are grouped into two families. The first family includes  $m_1$  endpoints that are adequately powered and the second family includes  $m_2$  potentially underpowered endpoints ( $m_1 + m_2 = m$ ). The endpoints in the first family are tested using any FWER controlling procedure at level  $\alpha_1 = \alpha - \varepsilon$ , where  $\varepsilon > 0$  is small, e.g.,  $\alpha = 0.05$  and  $\varepsilon = 0.005$ . For example, the Hochberg procedure decides that all endpoints in the first family are significant if  $p_{(m_1)} \leq \alpha_1$ , where  $p_{(m_1)}$  is the maximum  $p$ -value associated with those endpoints. The endpoints in the other family are tested using any FWER controlling procedure at level  $\alpha_2$ , which is *adaptively* based on  $p_{(m_1)}$  as follows:

$$\alpha_2(p_{(m_1)}) = \begin{cases} \alpha & \text{if } p_{(m_1)} \leq \alpha_1, \\ \min(\alpha^*/p_{(m_1)}^2, \alpha_1) & \text{if } p_{(m_1)} > \alpha_1, \end{cases}$$

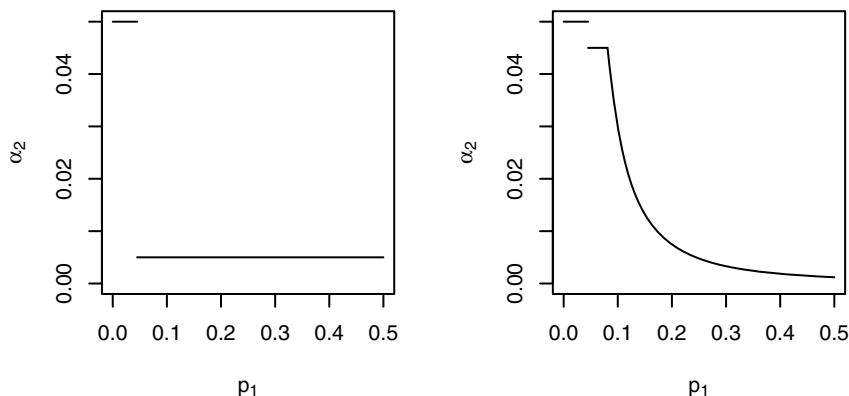
where

$$\alpha^* = \begin{cases} \alpha_1 \left(1 - \sqrt{2 - \alpha_1/m_1 - \alpha/\alpha_1}\right)^2 & \text{if } \alpha_1 + \alpha_1^2/m_1 - \alpha_1^3/m_1^2 \leq \alpha, \\ \alpha_1(\alpha - \alpha_1)/(m_1 - \alpha_1) & \text{if } \alpha_1 + \alpha_1^2/m_1 - \alpha_1^3/m_1^2 > \alpha. \end{cases}$$

It should be pointed out that this derivation assumes that all the  $p$ -values are independent. Li and Mehrotra also proposed an empirical adjustment to  $\alpha^*$  if the endpoints follow a multivariate normal distribution.

An advantage of this method over the Bonferroni-based PAAS method is that the remaining endpoints are tested at a generally higher significance level, which improves their power. As an example, consider the case of two independent endpoints (Endpoints 1 and 2) and let  $\alpha = 0.05$  and  $\alpha_1 = 0.045$ . The relationship between the significance level for Endpoint 2 ( $\alpha_2$ ) and  $p$ -value for Endpoint 1 ( $p_1$ ) is depicted in [Figure 4.1](#) (solid curve). The significance level for Endpoint 2 is 0.05 when  $p_1 \leq 0.045$ , 0.045 when  $0.045 \leq p_1 \leq 0.081$  and remains higher than 0.0052 (the significance level for the PAAS method) when  $p_1 \leq 0.244$ .

Conceptually, the 4A procedure is similar to the fallback procedure except that in the latter  $\alpha_2$  takes only two values depending on whether  $p_1 \leq \alpha_1$  or  $> \alpha_1$ . To compare the two procedures, consider again the clinical trial example with two independent endpoints. [Figure 4.1](#) depicts the significance level for Endpoint 2 as a function of the  $p$ -value for Endpoint 1 for the fallback and 4A procedures. The two procedures test Endpoint 1 at the same level if  $p_1 \leq 0.045$ . The significance level for Endpoint 2 for the 4A method is less



**FIGURE 4.1:** The significance level for Endpoint 2 ( $\alpha_2$ ) as a function of the  $p$ -value for Endpoint 1 ( $p_1$ ) for the fallback procedure (left panel) and 4A procedure (right panel).

stringent than that for the fallback procedure when  $0.045 \leq p_1 \leq 0.244$  and more stringent when  $p_1 > 0.244$ .

### 4.3.2 Parametric and resampling-based procedures

One of the limitations of procedures based on univariate  $p$ -values in the analysis of multiple endpoints is that they ignore the correlations among the endpoints. One can improve the power of these procedures by constructing parametric (normal theory) or resampling-based procedures that take correlations into account.

#### Bonferroni-type parametric procedure

Assume that the  $m$  endpoints follow a multivariate normal distribution and let  $t_i$  denote the  $t$ -statistic for testing the  $i$ th endpoint. The single-step parametric procedure is conceptually similar to the Bonferroni procedure in that the  $m$  hypotheses are tested simultaneously (i.e., in a single step). The global null hypothesis of no treatment effect is rejected if at least one test is significant, i.e., if  $t_{\max} = \max(t_1, \dots, t_m) \geq c$ , where  $c$  is a critical value computed from  $P\{t_{\max} < c\} = 1 - \alpha$ . This calculation is performed under the global hypothesis (which is the least favorable configuration at which the

Type I error probability of this procedure is maximized over the null space). In other words,  $c$  is the  $(1 - \alpha)$ -quantile of  $t_{\max}$  when  $\delta_1 = \dots = \delta_m = 0$ .

In general, this critical value is difficult to evaluate because the joint distribution of  $t_1, \dots, t_m$  (termed the *generalized multivariate  $t$  distribution*) is not known except for the case of two endpoints (Siddiqui, 1967; Gong, Pinheiro and DeMets, 2000). Note that this distribution is different from the standard multivariate  $t$ -distribution used in the Dunnett (1955) procedure because the denominator of each  $t_i$  uses a different error estimate,  $s_i$ . An additional complicating factor is that the joint distributions of both the numerators and denominators of the  $t_i$  statistics depend on the unknown correlation matrix.

## Fallback-type parametric procedure

The main difficulty in computing critical values of the Bonferroni-type parametric procedure is that the  $m$  test statistics are evaluated simultaneously and one has to deal with a complicated null distribution. A stepwise method that examines the endpoints in a sequential manner considerably simplifies the process of calculating the null distributions of the test statistics and critical values. As an illustration, we will consider the stepwise procedure for multiple endpoints proposed by Huque and Alosch (2008). This procedure is a parametric extension of the fallback procedure introduced in Section 4.3.1.

Unlike the regular fallback procedure, the parametric fallback procedure takes into account the joint distribution of the test statistics associated with individual endpoints, which leads to improved power for endpoints placed later in the sequence. As before, let  $t_1, \dots, t_m$  denote the test statistics for the  $m$  endpoints and let  $w_1, \dots, w_m$  denote weights that represent the importance of the endpoints (the weights are positive and add up to 1). The test statistics are assumed to follow a standard multivariate normal distribution.

The first step involves computation of critical values  $c_1, \dots, c_m$  and significance levels  $\gamma_1, \dots, \gamma_m$  that are defined recursively using the following equations:

$$\begin{aligned} P(t_1 \geq c_1) &= \alpha w_1, \\ P(t_1 < c_1, \dots, t_{i-1} < c_{i-1}, t_i \geq c_i) &= \alpha w_i, \quad i = 2, \dots, m. \end{aligned}$$

The probabilities are computed under the global null hypothesis. The significance levels associated with these critical values are defined as  $\gamma_i = 1 - \Phi(c_i)$ ,  $i = 1, \dots, m$ , where  $\Phi(x)$  is the cumulative distribution function of the standard normal distribution. Given these significance levels, the testing algorithm is set up as follows. The first endpoint is tested at the  $\gamma_1$  level (note that  $\gamma_1 = \alpha w_1$  and thus the parametric procedure uses the same level for the first endpoint as the regular procedure). At the  $i$ th step of the algorithm, the level is determined by the significance of the endpoints placed earlier in the sequence. For example, if  $s$  is the index of the last non-significant endpoint, the significance level for the  $i$ th endpoint is given by  $\max(\alpha w_{s+1} + \dots + \alpha w_i, \gamma_i)$ . In

**TABLE 4.2:** Power of the parametric fallback procedure in a clinical trial with three endpoints as a function of the effect sizes and correlation. The procedure is carried out at the one-sided 0.025 level using the (0.5, 0.25, 0.25) weighting scheme.

Correlation	Power of individual tests (%) (Endpoint 1, Endpoint 2, Endpoint 3)
Scenario 1 ( $e_1 = 0.4, e_2 = 0.4, e_3 = 0.4$ )	
0	(69.5, 72.4, 73.6)
0.2	(69.5, 71.9, 73.0)
0.5	(69.5, 71.5, 73.2)
Scenario 2 ( $e_1 = 0.3, e_2 = 0.4, e_3 = 0.4$ )	
0	(43.2, 68.6, 71.5)
0.2	(43.2, 67.9, 70.9)
0.5	(43.2, 67.6, 71.0)
Scenario 3 ( $e_1 = 0.5, e_2 = 0.4, e_3 = 0.4$ )	
0	(89.9, 74.9, 75.1)
0.2	(89.9, 74.9, 75.2)
0.5	(89.9, 74.9, 75.2)

other words, as in the regular procedure, the more consecutive endpoints are found significant, the higher the level for the current endpoint. However, if there is no evidence of a significant treatment effect for the  $(i - 1)$ th endpoint, the  $i$ th endpoint is tested at the  $\gamma_i$  level.

To help the reader appreciate the benefits of the parametric approach, consider the example from Wiens (2003) used in Section 4.3.1. This example deals with a clinical trial with two unequally weighted endpoints ( $w_1 = 0.8$  and  $w_2 = 0.2$ ) tested at the overall two-sided  $\alpha = 0.05$ . Using the regular procedure, the first endpoint is tested at  $\alpha w_1 = 0.04$  and, if the outcome is significant, the second endpoint is tested at the full 0.05 level. Otherwise, the significance level for the second endpoint is  $\alpha w_2 = 0.01$ . To apply the parametric procedure, assume that the test statistics for these two endpoints follow a standard bivariate normal distribution with the correlation coefficient  $\rho$ . The first endpoint is tested at  $\gamma_1 = 0.04$  and, if a significant result is observed, the other endpoint is tested at the 0.05 level. Thus these two levels are identical to those used by the fallback procedure above. However, if the first endpoint is not significant, the level for the second endpoint can be higher than  $\alpha w_2 = 0.01$ . The parametric procedure tests the second endpoint at 0.0104 for  $\rho = 0$ , 0.0112 for  $\rho = 0.3$  and 0.0146 for  $\rho = 0.6$  (see Table 1 in Huque and Alosch, 2008).

Table 4.2 summarizes the power of the parametric fallback procedure in the setting described in Section 4.3.1. The weights assigned to the three tests are  $w_1 = 0.5$ ,  $w_2 = 0.25$  and  $w_3 = 0.25$ .

The parametric procedure is uniformly more powerful than the regular procedure in all three scenarios but, in general, the two power functions are

quite close to each other (the difference is less than two percentage points). The parametric procedure exhibits the same key features as the regular procedure, e.g.,

- The parametric procedure is robust with respect to the monotonicity assumption and performs well when the first test in the sequence is underpowered.
- When the effect sizes across the tests are comparable, the power of individual tests improves toward the end of the sequence.
- The power of tests later in the sequence declines with increasing correlation.

Additional simulations performed by Huque and Alosch (2008) for the case of two hypotheses demonstrated that the power of the parametric procedure is comparable to that of the regular procedure when the test statistics are uncorrelated or effect sizes are equal regardless of the weighting scheme. The power advantage of the parametric procedure for the second test increases with the increasing correlation when the effect size of the second test is greater than that of the first test.

## Resampling-based procedures

Given the challenges associated with single-step parametric procedures for multiple endpoints (the joint distribution of the test statistics depends on an unknown correlation matrix), one can consider an alternative approach that uses the resampling-based methodology developed by Westfall and Young (1993). This alternative was explored by Reitmeir and Wassmer (1999) who introduced resampling-based versions of several single-step and stepwise procedures, e.g., Bonferroni and Hommel procedures, in the context of the multiple endpoint problem.

Along the lines of the general resampling-based method (see Section 2.8 for a detailed description of resampling-based procedures), Reitmeir and Wassmer proposed to estimate the joint distribution of the test statistics under the global null hypothesis using the bootstrap. Beginning with any multiple test, a resampling-based at-least-one procedure can be constructed using the following algorithm:

- Let  $p_i$  be the  $p$ -value for the  $i$ th endpoint,  $i = 1, \dots, m$  (this  $p$ -value is computed using a selected test).
- Generate  $K$  bootstrap samples (draw random samples with replacement of the same size as the original samples). Let  $p_i(k)$  be the treatment comparison  $p$ -value for the  $i$ th endpoint, which is computed from the  $k$ th bootstrap run using the selected test,  $i = 1, \dots, m$  and  $k = 1, \dots, K$ .



- Define the bootstrap multiplicity adjusted  $p$ -value for the  $i$ th endpoint as the proportion of bootstrap runs in which  $p_i(k) \leq p_i$ ,  $1, \dots, m$ .

The treatment effect for the  $i$ th endpoint is significant if the bootstrap multiplicity adjusted  $p$ -value is no greater than the pre-specified familywise error rate  $\alpha$ . Reitmeir and Wassmer showed via simulations that the resampling-based tests for the multiple endpoint problem resulted in a consistent power gain compared to the original tests. The improvement in power was rather small for the Bonferroni test; however, a substantially larger gain was observed for some other tests, e.g., the Hommel test.

## 4.4 Global testing procedures

An important property of global testing procedures is that they combine evidence of treatment effect across several endpoints and thus they are more powerful than procedures for individual endpoints (provided the treatment effects are consistent across the endpoints). In this section we first consider procedures for the goal of demonstrating overall efficacy of the treatment and then describe inferences for individual endpoints when the global assessment produces a significant result.

### 4.4.1 Normal theory model

Global testing procedures considered in this section (with a few exceptions such as the global rank-sum procedure introduced by O'Brien, 1984) assume the normal theory model described below.

Consider a two-arm clinical trial with a parallel-group design in which the treatment group (Group 1) is tested versus the control group (Group 2). As in Lehmacher, Wassmer and Reitmeir (1991), the response of the  $j$ th patient in the  $i$ th group with respect to the  $k$ th endpoint is denoted by  $X_{ijk}$ ,  $i = 1, 2$ ,  $j = 1, \dots, n_i$ ,  $k = 1, \dots, m$ . Let

$$\bar{X}_{i \cdot k} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ijk} \quad (i = 1, 2, 1 \leq k \leq m).$$

The vector of patient responses on the  $m$  endpoints,  $(X_{ij1}, \dots, X_{ijm})$ , is assumed to follow a multivariate normal distribution with a mean vector  $(\mu_{i1}, \dots, \mu_{im})$  and a common covariance matrix  $\Sigma$ . The diagonal elements of the covariance matrix are  $\sigma_k^2 = \text{var}(X_{ijk})$ ,  $k = 1, \dots, m$ . The correlation matrix of the endpoints is denoted by  $R$  and its elements by  $\rho_{kl}$ . One may think that multiple endpoints are always highly correlated. In fact, independent endpoints are desirable because they are not proxies of each other and

thus contain more information. The correlations  $\rho_{kl}$  rarely exceed 0.6 in clinical trial applications; see Sankoh, D'Agostino and Huque (2003).

The mean treatment difference for the  $k$ th endpoint is defined as  $\delta_k = \mu_{1k} - \mu_{2k}$  and it is assumed that large values of  $\delta_k$  imply higher treatment efficacy. To define the test statistic for the  $k$ th endpoint, let  $\bar{X}_{i \cdot k}$  denote the mean response in the  $i$ th group on the  $k$ th endpoint and let  $S$  denote the pooled sample covariance matrix. The diagonal elements of  $S$  are denoted by  $s_1^2, \dots, s_m^2$ . The treatment effect for the  $k$ th endpoint is tested using the  $t$ -statistic

$$t_k = \frac{\bar{X}_{1 \cdot k} - \bar{X}_{2 \cdot k}}{s_k \sqrt{1/n_1 + 1/n_2}}.$$

#### 4.4.2 OLS and GLS procedures

To motivate procedures described in this section, consider the Bonferroni procedure for multiple endpoints from Section 4.3.1. The global version of this procedure rejects the hypothesis  $H_I$  from (4.1) if

$$p_{\min} = \min(p_1, \dots, p_m) \leq \alpha/m.$$

This procedure depends only on the smallest  $p$ -value and ignores all other  $p$ -values. Therefore it is not sensitive in the common scenario where small to modest effects are present in all endpoints.

To address this shortcoming of the Bonferroni global procedure, O'Brien (1984) considered the setup in which the multivariate testing problem is simplified by making an assumption of a *common standardized effect size*. Specifically, assume that the standardized effect sizes for the  $m$  endpoints,  $\delta_1/\sigma_1, \dots, \delta_m/\sigma_m$ , are equal to, say,  $\lambda$ . In this case, the problem of testing the null hypothesis,

$$H_I^* : \delta_i = 0 \text{ for all } i,$$

reduces to a single parameter testing problem

$$H^* : \lambda = 0 \text{ versus } K^* : \lambda > 0.$$

#### OLS and GLS test statistics

O'Brien (1984) proposed two procedures for the hypothesis  $H^*$  based on standardized responses,  $Y_{ijk} = X_{ijk}/\sigma_k$ . Under the simplifying assumption of a common effect size, one can consider the following regression model for the standardized responses:

$$Y_{ijk} = \frac{\mu_k}{\sigma_k} + \frac{\lambda}{2} I_i + e_{ijk},$$

where  $i = 1, 2$ ,  $j = 1, \dots, n_i$ ,  $k = 1, \dots, m$ ,  $\mu_k = (\mu_{1k} + \mu_{2k})/2$ ,  $I_i = +1$  if  $i = 1$  and  $-1$  if  $i = 2$ , and  $e_{ijk}$  is  $N(0, 1)$  distributed error term with  $\text{corr}(e_{ijk}, e_{i'j'k'}) = \rho_{kk'}$  if  $i = i'$  and  $j = j'$ , and 0 otherwise.

The first procedure developed by O'Brien is based on the ordinary least squares (OLS) estimate of the common effect size  $\lambda$  while the second procedure is based on the generalized least squares (GLS) estimate. Let  $\hat{\lambda}_{\text{OLS}}$  and  $\text{SE}(\hat{\lambda}_{\text{OLS}})$  denote the OLS estimate of  $\lambda$  and its sample standard error, respectively. It can be shown that the OLS test statistic for  $H^*$  is given by

$$t_{\text{OLS}} = \frac{\hat{\lambda}_{\text{OLS}}}{\text{SE}(\hat{\lambda}_{\text{OLS}})} = \frac{J't}{\sqrt{J'\hat{R}J}},$$

where  $J$  is an  $m$ -vector of all 1's and  $t = (t_1, \dots, t_m)'$  is the vector of  $t$ -statistics defined in Section 4.4.1.

Since the error terms  $e_{ijk}$  in the regression model for the standardized responses are correlated, it may be preferable to use the GLS estimate of  $\lambda$ , which leads to the following test statistic for  $H^*$

$$t_{\text{GLS}} = \frac{\hat{\lambda}_{\text{GLS}}}{\text{SE}(\hat{\lambda}_{\text{GLS}})} = \frac{J'\hat{R}^{-1}t}{\sqrt{J'\hat{R}^{-1}J}}.$$

It is instructive to compare the OLS and GLS test statistics. Both testing procedures assess the composite effect of multiple endpoints by aggregating the  $t$ -statistics for the individual endpoints. In the case of the OLS test statistic, the  $t$ -statistics are equally weighted, while the GLS test statistic assigns unequal weights to the  $t$ -statistics. The weights are determined by the sample correlation matrix  $\hat{R}$ . If a certain endpoint is highly correlated with the others, it is not very informative, so the GLS procedure gives its  $t$ -statistic a correspondingly low weight. A downside of this approach is that the weights can become negative. This leads to anomalous results, e.g., it becomes possible to reject  $H^*$  even if the treatment effect is negative on all the endpoints.

In order to compute critical values for the OLS and GLS procedures, one needs to derive the null distributions of their test statistics. For large sample sizes, the OLS and GLS statistics approach the standard normal distribution under  $H^*$ , but the approach of the GLS statistic is slower since it has the random matrix  $\hat{R}$  both in the numerator and denominator. For small sample sizes the standard normal distribution provides a liberal test of  $H^*$ . The exact small sample null distributions of  $t_{\text{OLS}}$  and  $t_{\text{GLS}}$  are not known. O'Brien (1984) proposed a  $t$ -distribution with  $\nu = n_1 + n_2 - 2m$  df as an approximation. This approximation is exact for  $m = 1$  but conservative for  $m > 1$ . Logan and Tamhane (2004) proposed the approximation,  $\nu = 0.5(n_1 + n_2 - 2)(1 + 1/m^2)$ , which is more accurate.

Logan and Tamhane (2004) also extended the OLS and GLS procedures to the heteroscedastic case (case of unequal  $\Sigma$ s). Note that the heteroscedastic extension of the GLS test statistic given in Pocock, Geller and Tsiatis (1987) does not have the standard normal distribution under  $H^*$  as claimed there, and hence should not be used.

**TABLE 4.3:** Summary of the results of the osteoarthritis trial (SD, standard deviation).

Endpoint	Summary statistic	Treatment <i>n</i> = 88	Placebo <i>n</i> = 90
Pain subscale	Mean Pooled SD	59 96	35 96
Physical function subscale	Mean Pooled SD	202 278	111 278

**Osteoarthritis trial example**

To illustrate the use of global testing procedures, consider a clinical trial for the treatment of osteoarthritis. The study was conducted to evaluate the effects of a treatment on two endpoints, the pain and physical function subscales of the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index (Bellamy, 2002), compared to placebo. The efficacy analysis was based on the mean changes in the two endpoints during a 6-week study period. The results of the study are shown in Table 4.3.

The OLS procedure was carried out to assess the overall efficacy of the treatment (note that the GLS procedure is equivalent to the OLS procedure in the case of two endpoints). The *t*-statistics for the pain and physical function endpoints and sample correlation matrix were

$$\begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \begin{bmatrix} 1.67 \\ 2.18 \end{bmatrix}, \quad \widehat{R} = \begin{bmatrix} 1 & 0.36 \\ 0.36 & 1 \end{bmatrix}.$$

Given this information, it is easy to compute the OLS/GLS test statistic,

$$t_{\text{OLS}} = t_{\text{GLS}} = \frac{1.67 + 2.18}{\sqrt{1 + 0.36 + 0.36 + 1}} = 2.334.$$

Using the Logan-Tamhane formula, the one-sided *p*-value associated with this test statistic is 0.0107 (based on 110 df), which is significant at the one-sided 0.025 level.

It is worth noting that the OLS procedure becomes less powerful as the correlation between two endpoints increases. In fact, the OLS statistic would not be significant at the one-sided 0.025 level in the osteoarthritis trial if the sample correlation coefficient was greater than 0.91. The reason is that higher correlations imply correspondingly less independent information in the endpoints.

**Power calculations for OLS and GLS procedures**

From their construction, it is clear that the OLS and GLS procedures will be powerful when all endpoints have a similar positive effect, but in other

situations, they may lack power. Dallow, Leonov and Roger (2008) considered the problem of power and sample size calculations for the OLS and GLS procedures and introduced a simple measure, termed the *operational effect size*, that helps to quantify the composite effect of multiple endpoints. Let  $\lambda_k = \delta_k/\sigma_k$  be the true standardized effect size for the  $k$ th endpoint and let  $\lambda = (\lambda_1, \dots, \lambda_m)'$ . Under the alternative hypothesis, the distributions of  $t_{\text{OLS}}$  and  $t_{\text{GLS}}$  can be approximated by noncentral  $t$ -distributions with noncentrality parameters given by

$$\Delta_{\text{OLS}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{J' \lambda}{\sqrt{J' R J}} \text{ and } \Delta_{\text{GLS}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{J' R^{-1} \lambda}{\sqrt{J' R^{-1} J}}.$$

Given this, Dallow et al. defined the operational effect sizes of the two procedures as

$$\Lambda_{\text{OLS}} = \frac{J' \lambda}{\sqrt{J' R J}} \text{ and } \Lambda_{\text{GLS}} = \frac{J' R^{-1} \lambda}{\sqrt{J' R^{-1} J}}.$$

The two quantities serve the same role in the problem of testing multiple endpoints as the regular effect size in a single-endpoint problem. It is easy to show that  $\Lambda_{\text{OLS}}$  and  $\Lambda_{\text{GLS}}$  are in fact the standardized effect sizes for an equivalent single endpoint which would have the same overall power for the same sample sizes,  $n_1$  and  $n_2$ .

Operational effect sizes help establish the relationship between key properties of multiple endpoints (e.g.,  $\lambda$  and  $R$ ) and the power of the two global procedures. For example, the numerator in  $\Lambda_{\text{OLS}}$  is the sum of the effect sizes for individual endpoints and thus the power of the OLS procedure is an increasing function of each  $\lambda_k$ . Further, the denominator depends on the correlations among the endpoints. It is clear that, as the endpoints become more correlated and thus less informative, the operational effect size decreases in magnitude. In addition, operational effect sizes facilitate power comparisons between the OLS and GLS procedures. Dallow et al. proved that the GLS procedure is more powerful than the OLS procedure when the effect sizes are equal across the endpoints ( $\lambda_1 = \dots = \lambda_m$ ) but, in general, the power of the GLS procedure can be lower than that of the OLS procedure. This happens, for example, when the effect sizes of strongly correlated endpoints are large and the effect sizes of weakly correlated endpoints are small.

Several authors have reported results of simulation studies to assess the power of the OLS and GLS procedures under various configurations of effect sizes and correlation values. The simulation study by Reitmeir and Wassmer (1996) showed that the power of the OLS procedure was comparable to that of the GLS procedure. Dallow et al. (2008) demonstrated that the GLS procedure is biased when the sample size is small to moderate which complicates the power comparison of the OLS and GLS procedures for  $n \leq 40$ . For larger sample sizes, the difference between the power functions of the two procedures is very small.

Power and sample size calculations can be performed using a normal approximation by specifying  $\lambda$  and  $R$ . Suppose, for example, that we are inter-

ested in computing the number of patients in each arm of a two-arm trial. The one-sided Type I error rate is given by  $\alpha$  and the desired power of a global procedure (OLS or GLS) is set at  $1 - \beta$ . As shown by Dallow et al., the sample size in each arm is given by the familiar formula,

$$n = \frac{2(z_\alpha + z_\beta)^2}{\Lambda^2},$$

where  $z_x$  is the  $(1 - x)$ -quantile of the standard normal distribution and  $\Lambda$  is the operational effect size of the global procedure chosen in this trial.

As an illustration, we will return to the osteoarthritis clinical trial example and compute the sample size required to achieve 90% power at a one-sided 0.025 level ( $\alpha = 0.025$  and  $\beta = 0.1$ ). Based on the results displayed in [Table 4.3](#), assume that the standardized effect sizes for the pain and physical function endpoints are 0.25 and 0.33, respectively, and the correlation coefficient is 0.36. Under these assumptions, the operational effect size for the OLS and GLS procedures is

$$\Lambda_{\text{OLS}} = \Lambda_{\text{GLS}} = \frac{0.25 + 0.33}{\sqrt{1 + 0.36 + 0.36 + 1}} = 0.35$$

and thus

$$n = \frac{2(1.96 + 1.28)^2}{0.35^2} \simeq 170$$

patients per arm need to be enrolled in the study.

Given that the GLS procedure does not generally dominate the OLS procedure in terms of power and because of the added difficulties caused by negative weights in the GLS statistic, we recommend the use of the OLS procedure in clinical trials.

## Nonparametric global procedures

The OLS and GLS procedures can be formulated for non-normal responses as long as the test statistics for the  $m$  endpoints follow a multivariate normal distribution in large samples. Examples include binary and time-to-event variables (Pocock, Geller and Tsiatis, 1987). However, if the assumption of multivariate normality cannot be made, one can consider a nonparametric version of the OLS procedure proposed by O'Brien (1984). In this procedure, the data from the two groups are pooled and ranked on each endpoint separately as in the Wilcoxon rank-sum test. Let  $r_{ijk}$  be the rank of  $X_{ijk}$  in the pooled sample. Then a two-sample  $t$ -test is performed on the summed ranks,

$$r_{ij} = \sum_{k=1}^m r_{ijk}, \quad i = 1, 2, \quad j = 1, \dots, n_i.$$

This procedure offers a viable alternative to the OLS procedure particularly if the data are non-normal. For example, this global rank-sum procedure was

used in the azithromycin study in patients with coronary artery disease (Anderson et al., 1999). The procedure was chosen to evaluate the overall effect of the treatment on four inflammatory markers because the change scores were not expected to follow a normal distribution.

#### 4.4.3 Likelihood ratio and other procedures

This section gives a review of a class of global procedures based on the likelihood ratio (LR) principle and two other global procedures (the L  uter and Follmann procedures). This review is rather brief because these procedures are not commonly used in clinical trial applications due to the limitations discussed below.

#### Exact likelihood ratio procedures

It will be assumed in this section that the point null and one-sided alternative hypotheses are given by (4.2). Kud   (1963) was the first to derive an exact one-sided LR procedure for the one-sample problem when the covariance matrix  $\Sigma$  is known. This procedure can be readily extended to the two-sample problem. Perlman (1969) extended the Kud   procedure to the case of an unknown covariance matrix but the null distribution of the resulting test statistic is not free of  $\Sigma$  and the procedure is biased. However, Perlman provided sharp lower and upper bounds on the null distribution that are free of  $\Sigma$ . Wang and McDermott (1998) solved the problem of dependence on unknown  $\Sigma$  by deriving an LR procedure conditional on the sample covariance matrix  $S$ .

These procedures are not commonly used because they are not easy to implement computationally. There is, however, a more basic problem with the LR procedures that they can reject  $H_I^*$  even when the vector of mean treatment differences has all negative elements (Silvapulle, 1997). These procedures are also nonmonotone in the sense that if the differences  $\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k}$  become more negative the test statistic can get larger.

Perlman and Wu (2002) showed that these difficulties are caused by the point null hypothesis. Basically, the LR procedure compares the *ratio* of the likelihood under  $K_U^*$  versus that under  $H_I^*$ . The apparent nonmonotonicity of the LR procedure results because, in some cases, as the sample outcomes move deeper into the part of the sample space corresponding to  $K_U^*$ , their likelihood under  $K_U^*$  increases, but so does their likelihood under  $H_I^*$ , and their ratio gets smaller. This is not a defect of the LR procedure, but rather that of the null hypothesis not being correctly specified. If the null hypothesis is defined as a full complement of  $K_U^*$  then the LR procedure no longer has these difficulties. However, computation of the test statistic under the complete null hypothesis and its null distribution are problematic.

Approximate LR procedures were proposed in the literature to circumvent the computational and analytical difficulties of the exact LR procedure; see, for example, Tang, Gnecco and Geller (1989) and Tamhane and Logan (2002).

These procedures, although easier to apply, suffer from the same anomalies that the exact LR procedures suffer because of the misspecification of the null hypothesis as a point null hypothesis.

Cohen and Sackrowitz (1998) proposed the cone-ordered monotone (COM) criterion to overcome the nonmonotonicity problem. However, their COM procedure is not entirely satisfactory either since, e.g., in the bivariate case, it can reject  $H_I^*$  if one mean difference is highly negative as long as the other mean difference is highly positive.

### Läuter exact procedure

Läuter (1996) and Läuter, Glimm and Kropf (1996) proposed a class of test statistics having the property that they are exactly  $t$ -distributed under the point null hypothesis. To define the procedure, let  $\bar{X}_{..k}$  denote the overall sample mean for the  $k$ th endpoint, i.e.,

$$\bar{X}_{..k} = \frac{1}{n_1 + n_2} \sum_{i=1}^2 \sum_{j=1}^{n_i} X_{ijk}.$$

Consider the total cross-products matrix  $V$  with elements

$$v_{kl} = \sum_{i=1}^2 \sum_{j=1}^{n_i} (X_{ijk} - \bar{X}_{..k})(X_{ijl} - \bar{X}_{..l}), \quad k, l = 1, \dots, m.$$

Let  $w = w(V)$  be any  $m$ -dimensional vector of weights depending only on  $V$  and  $w \neq 0$  with probability 1. Läuter (1996) showed that

$$t_w = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left( \frac{w' t}{\sqrt{w' S w}} \right)$$

is  $t$ -distributed with  $n_1 + n_2 - 2$  df under the point null hypothesis. Various choices of  $w$  were discussed by Läuter et al. (1996). The simplest among them is  $w_k = 1/\sqrt{v_{kk}}$ . The resulting statistic is called the standardized sum (SS) statistic (denoted by  $t_{SS}$ ) which can be expressed as

$$t_{SS} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left( \frac{\bar{Y}_{1.} - \bar{Y}_{2.}}{s_y} \right),$$

where  $\bar{Y}_{i.}$  is the average of the standardized observations

$$Y_{ij} = \sum_{k=1}^m \frac{X_{ijk}}{\sqrt{v_{kk}}}, \quad i = 1, 2, \quad j = 1, \dots, n_i.$$

Analytical and simulated power comparisons made in Logan and Tamhane (2004) showed that the OLS procedure is more powerful than the Läuter procedure when only a few endpoints have an effect. In fact, if only one endpoint



has an effect, which tends to infinity, the power of the Lauter procedure remains bounded strictly below 1, whereas the power of the OLS procedure tends to 1, as it should. Frick (1996) also noted this drawback of the Lauter procedures but argued that such a scenario is unlikely in practice. When all endpoints have roughly equal effects, the powers of the OLS and Lauter procedures are comparable.

The reason for the lack of power of the SS test statistic when only a few endpoints have an effect is that it standardizes the data on each endpoint by its pooled *total* group sample standard deviation and then computes an overall *t*-statistic. The pooled standard deviation overestimates the true standard deviation since it includes the differences between the means of the treatment and control groups which diminishes the power of the Lauter procedure. On the other hand, the OLS statistic is the sum of *t*-statistics obtained by standardizing the individual endpoints by their pooled *within* group sample standard deviations.

## Follmann procedure

Follmann (1996) proposed an ad-hoc procedure which is simple to apply: Reject the point null hypothesis if the two-sided Hotelling's  $T^2$ -test is significant at the  $2\alpha$ -level and the average endpoint mean difference is positive,

$$\sum_{k=1}^m (\bar{X}_{1\cdot k} - \bar{X}_{2\cdot k}) > 0.$$

Unfortunately, the alternative for which this procedure is designed,

$$\sum_{k=1}^m (\mu_{1k} - \mu_{2k}) > 0,$$

is not very meaningful since it depends on the scaling used for the endpoints.

### 4.4.4 Procedures for individual endpoints

As was explained earlier in this section, global procedures are aimed at an overall evaluation of the treatment effect. However, if the overall treatment effect is positive, the trial's sponsor is likely to be interested in examining the treatment effect on individual endpoints/components. As pointed out in Chapter 1, the LIFE trial (Dahlof et al., 2002) serves as an example of a trial in which the analysis of individual endpoints provided important insights into the nature of treatment benefit. This analysis revealed that the overall treatment effect was driven mainly by one component (stroke endpoint). Here we review extensions of global testing methods that can be used to perform multiplicity-adjusted univariate inferences for individual endpoints after a significant global result.

Lehmacher, Wassmer and Reitmeir (1991) applied the closure principle to make inferences on individual endpoints. As was explained in Section 2.3.3, this principle is a popular tool used in the construction of multiple testing procedures. For example, the Holm procedure is a closed procedure for testing individual hypotheses which is derived from the Bonferroni global procedure (see [Section 4.4.2](#)). A similar approach can be used to construct a closed testing procedure for testing individual endpoints based on any other global procedure. To accomplish this, one needs to consider all possible combinations of the  $m$  endpoints and test each combination using an  $\alpha$ -level global procedure subject to the closure principle, i.e., if the procedure for any combination is not significant then all of its subset combinations are declared nonsignificant without testing them. The treatment effect on an endpoint is significant at level  $\alpha$  if the global procedures for all the combinations including the selected endpoint are significant at this level.

This approach can be used with any  $\alpha$ -level global procedure for testing different intersections. Lehmacher et al. constructed a procedure for testing individual endpoints based on the OLS and GLS procedures described in Section 4.4.2. Wang (1998) applied the Follmann procedure introduced in Section 4.4.3 as the global procedure and found the performance comparable to the Westfall-Young resampling procedures (Section 2.8). Logan and Tamhane (2001) proposed a hybrid approach that uses a combination of global procedures, each one of which is powerful against a different alternative. Specifically, the Logan-Tamhane hybrid procedure consists of the Bonferroni global procedure based on the smallest  $p$ -value and the OLS procedure. Here the former procedure is powerful against alternatives where only a few endpoints have large effects while the latter procedure is powerful against alternatives where all endpoints have small to modest effects. The hybrid test statistic for each intersection hypothesis is the minimum of the  $p$ -values for the Bonferroni and OLS procedures. A bootstrap method is used to estimate the adjusted  $p$ -value for this complex statistic. The resulting procedure has stable and high power against a range of alternatives (i.e., the procedure is robust), but is computationally more intensive.

## Osteoarthritis trial example

A clinical study with two endpoints (pain and physical function) was considered in the osteoarthritis trial example (Section 4.4.2). The overall treatment effect of the two endpoints was evaluated using the OLS procedure. The global procedure was significant at the one-sided 0.025 level. We will apply the closure principle to assess the treatment effect on each endpoint and see whether the overall positive result was driven by both components or only one. First, we need to compute  $p$ -values for the endpoint-specific tests. From Section 4.4.2, the  $t$ -statistics for the pain and physical function endpoints are 1.67 and 2.18, respectively, with  $n_1 + n_2 - 2 = 172$  df. The one-sided  $p$ -values associated with the  $t$ -statistics are 0.0486 and 0.0152 applying the

Logan-Tamhane formula for degrees of freedom. Using the closure principle, the multiplicity adjusted  $p$ -value for each endpoint is the larger of the  $p$ -value for the OLS procedure and the  $p$ -value for the endpoint-specific  $t$ -test. The treatment's effect on the pain endpoint is not significant at the one-sided 0.025 level ( $p = 0.0486$ ) whereas the effect on the other endpoint is significant ( $p = 0.0152$ ). It is worth remembering that it is possible for endpoint-specific tests to be non-significant even if the global procedure is highly significant.

## 4.5 All-or-none procedures

As was explained in Section 4.2.3, the goal of demonstrating the efficacy of the treatment on *all* endpoints requires an all-or-none or IU procedure (Berger, 1982) of the union of individual hypotheses,  $H_i$ . The all-or-none procedure has the following form:

$$\text{Reject all hypotheses if } t_{\min} = \min_{1 \leq i \leq m} t_i \geq t_{\alpha}(\nu),$$

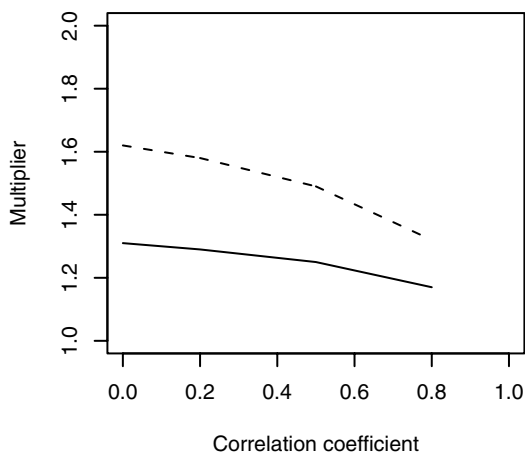
where  $t_{\alpha}(\nu)$  is the  $(1 - \alpha)$ -quantile of the  $t$ -distribution with  $\nu = n_1 + n_2 - 2$  df. This procedure is popularly known as the min test (Laska and Meisner, 1989).

Since this procedure does not use a multiplicity adjustment (each hypothesis  $H_i$  is tested at level  $\alpha$ ), it may appear at first that it must be highly powerful as a test of the global hypothesis  $H_U$ . In reality, the min test is very conservative because of the requirement that *all* hypotheses must be rejected at level  $\alpha$ . The conservatism results from the least favorable configuration of the min test which can be shown to be of the following form:

- No treatment effect for any one endpoint ( $\delta_i = 0$  for some  $i$ ).
- Infinitely large treatment effects for all other endpoints ( $\delta_j \rightarrow \infty$  for  $j \neq i$ ).

This configuration leads to marginal  $\alpha$ -level  $t$ -tests.

Figure 4.2, based on Offen et al. (2007), will help the reader appreciate how conservative the min test can be. This figure gives the multipliers to calculate the sample size required to guarantee 80% power for the min test if the base sample size guarantees 80% power for each endpoint. The calculation was done under the assumption of equicorrelated endpoints and a common standardized effect size for all endpoints. As can be seen from the table, the multiplier increases as the number of endpoints increases and the assumed common correlation between them decreases. Consider, for example, the clinical trial in patients with Alzheimer's disease described by Offen and Helderbrand (2003). It is commonly required that a new treatment should demonstrate a significant



**FIGURE 4.2:** Sample size multipliers for the min test for two endpoints (solid curve) and four endpoints (dashed curve) as a function of the common correlation in all-or-none testing problems. Multiplier equals 1 for a single endpoint.

effect on at least two endpoints, e.g., a cognition endpoint (Alzheimer's Disease Assessment Scale-Cognitive Subpart) and a clinical global scale (Clinician's Interview-Based Impression of Change). The correlation between these two endpoints is usually around 0.2 and thus the sample size multiplier is 1.29 which corresponds to almost a 30% increase in the sample size. In other cases, e.g., when four weakly correlated endpoints are considered, the multiplier is 1.58 meaning that the sample size needs to be increased by almost 60% compared to the single-endpoint case.

It is important to note that the least favorable configuration for the min test is clinically not plausible. The global hypothesis  $H_U$  permits configurations with infinitely large positive effects on some endpoints and negative effects on others. However, it is uncommon for treatments to have substantially different effects on the endpoints. This has led researchers to put restrictions on the global hypothesis in order to develop more powerful versions of the min test.

Hochberg and Mosier (2001) suggested restricting  $H_U$  to the negative quadrant,

$$\bigcap_{k=1}^m (\delta_k \leq 0),$$

in which case the least favorable configuration is the overall null configuration,  $\delta_1 = \dots = \delta_m = 0$ . Chuang-Stein et al. (2007) restricted the hypothesis to the subset of the global hypothesis which satisfies

$$\bigcap_{k=1}^m (-\varepsilon_k \leq \delta_k \leq \varepsilon_k),$$

where the thresholds  $\varepsilon_k$ ,  $k = 1, \dots, m$ , are prespecified based on clinical considerations. A similar approach, but based on estimated mean differences,  $\bar{X}_{1,k} - \bar{X}_{2,k}$ , was proposed by Snapinn (1987). Cappizi and Zhang (1996) suggested another alternative to the min test which requires that the treatment be shown effective at a more stringent significance level  $\alpha_1$  on say  $m_1 < m$  endpoints and at a less stringent significance level  $\alpha_2 > \alpha_1$  on the remaining  $m_2 = m - m_1$  endpoints. For  $m = 2$ , they suggested this rule for  $m_1 = m_2 = 1$ ,  $\alpha_1 = 0.05$  and  $\alpha_2 = 0.10$  or  $0.20$ . However, as pointed out by Neuhäuser, Steinijans and Bretz (1999), this rule does not control the FWER at  $\alpha = 0.05$ .

Another approach to this formulation adopts a modified definition of the error rate to improve the power of the min test in clinical trials with several endpoints. Chuang-Stein et al. (2007) considered an error rate definition based on the average Type I error rate over the null space and developed a procedure that adjusts significance levels for the individual endpoints to control the average Type I error rate at a prespecified level.

This is a relatively new research area and further work is required to assess the utility of the methods described in this section and their applicability.

## 4.6 Superiority-noninferiority procedures

There are many situations in which the requirement that the treatment be superior to the control on all endpoints (all-or-none procedures in Section 4.5) is often too strong and the requirement that the treatment be superior to the control on at least one endpoint (at-least-one procedures in Section 4.3) is too weak. The superiority-noninferiority approach discussed in this section strengthens the latter requirement by augmenting it with the additional requirement that the treatment is not inferior to the control on all other endpoints.

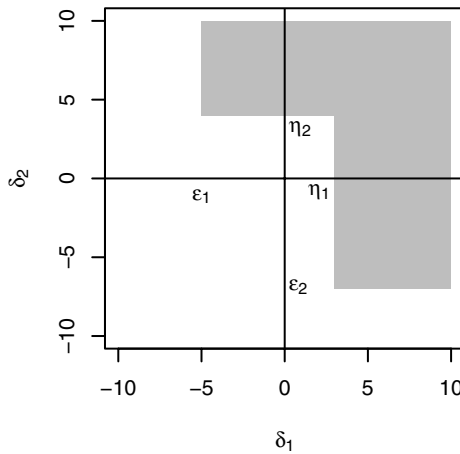
Consider a clinical trial with  $m$  endpoints and suppose that its objective is to demonstrate the treatment is superior to the control on at least one endpoint and noninferior to the control on all other endpoints. Note that if superiority is established for both endpoints, the second part of this requirement (noninferiority) becomes redundant. This formulation of the superiority-noninferiority testing problem was considered by Bloch, Lai and Tubert-Bitter (2001) and Tamhane and Logan (2004). The null and alternative hypotheses

for the superiority-noninferiority problem are defined in Section 4.2.4:

$$H_U^{(SN)} = H_I^{(S)} \cup H_U^{(N)} \text{ versus } K_I^{(SN)} = K_U^{(S)} \cap K_I^{(N)}.$$

The trial's outcome is declared positive if there is evidence of superior efficacy for at least one endpoint ( $K_U^{(S)}$ ) and noninferior efficacy for all endpoints ( $K_I^{(N)}$ ).

As an illustration, consider a clinical trial with two endpoints. The region corresponding to the alternative hypothesis  $K_I^{(SN)}$  with  $\varepsilon_1 = 5$ ,  $\varepsilon_2 = 7$ ,  $\eta_1 = 3$  and  $\eta_2 = 4$  is shown in Figure 4.3.



**FIGURE 4.3:** Region in the parameter space corresponding to the alternative hypothesis (shaded area) in a superiority-noninferiority testing problem with two endpoints.

### Tamhane-Logan superiority-noninferiority procedure

Denote the  $t$ -statistics for superiority and noninferiority for the  $k$ th endpoint by

$$t_k^{(S)} = \frac{\bar{X}_{1 \cdot k} - \bar{X}_{2 \cdot k} - \eta_k}{s_k \sqrt{1/n_1 + 1/n_2}}, \quad t_k^{(N)} = \frac{\bar{X}_{1 \cdot k} - \bar{X}_{2 \cdot k} + \varepsilon_k}{s_k \sqrt{1/n_1 + 1/n_2}}.$$

Tamhane and Logan (2004) used the UI statistic

$$t_{\max}^{(S)} = \max(t_1^{(S)}, \dots, t_m^{(S)})$$

for testing the superiority null hypothesis  $H_I^{(S)}$  and the IU statistic

$$t_{\min}^{(N)} = \min(t_1^{(N)}, \dots, t_m^{(N)})$$

for testing the noninferiority null hypothesis  $H_U^{(N)}$ . They proposed the following procedure of the global superiority-noninferiority hypothesis:

$$\text{Reject } H_U^{(SN)} \text{ if } t_{\max}^{(S)} \geq c^{(S)} \text{ and } t_{\min}^{(N)} \geq c^{(N)},$$

where the critical values  $c^{(S)}$  and  $c^{(N)}$  are chosen so that the procedure has level  $\alpha$ . Bloch et al. (2001) used the Hotelling  $T^2$ -statistic for testing superiority in place of  $t_{\max}^{(S)}$ ; however, as noted before, the  $T^2$ -statistic is not very powerful against one-sided superiority alternative. Perlman and Wu (2004) used Perlman's one-sided LR statistic instead of the  $T^2$ -statistic.

According to the intersection-union testing principle, the superiority and noninferiority tests must be of level  $\alpha$ . Conservative values for  $c^{(S)}$  and  $c^{(N)}$  can be chosen to be the  $(1 - \alpha/m)$ - and  $(1 - \alpha)$ -quantiles of the  $t$ -distribution, respectively, with  $\nu = n_1 + n_2 - 2$  df. The exact value of  $c^{(S)}$  involves the generalized multivariate  $t$  distribution and thus, as was explained in Section 4.3.2, is difficult to evaluate. Also, a sharper critical constant  $c^{(S)}$  can be evaluated by conditioning on the event that the noninferiority test is passed by all endpoints. However, the resulting value of  $c^{(S)}$  needs to be evaluated by using bootstrap; see Bloch et al. (2001) and Tamhane and Logan (2004). Röhmle et al. (2006) objected to this conditioning arguing that it causes significance of the superiority test to be influenced by changes in the noninferiority margin for which there is no clinical justification. However, Logan and Tamhane (2008) showed that passing the noninferiority test at a more stringent margin adds more credence to the alternative hypothesis  $K_U^{(S)}$  that the treatment is superior on at least one endpoint, and therefore it is retained more easily. If  $H_U^{(SN)} = H_I^{(S)} \cup H_U^{(N)}$  is rejected then it is of interest to know which endpoints demonstrate superiority of the treatment over the control. Logan and Tamhane (2008) gave a closed procedure for this purpose which controls FWER for the family of  $H_U^{(SN)}$  as well as the endpoint-specific superiority null hypotheses  $H_k^{(S)}$ . This closed procedure can be implemented in  $m + 1$  steps.

## Alzheimer's disease trial example

The Alzheimer's disease example from Section 4.5 will be used to illustrate key properties of the Tamhane-Logan superiority-noninferiority procedure. Table 4.4 displays results of a 24-week study in patients with Alzheimer's disease that tested the efficacy and safety of an experimental treatment compared to placebo. The efficacy profile of the treatment was evaluated using two co-primary endpoints, Alzheimer's Disease Assessment Scale-Cognitive

**TABLE 4.4:** Summary of the results of Alzheimer’s disease trial (SD, standard deviation).

Endpoint	Summary statistic	Treatment $n = 167$	Placebo $n = 161$
ADAS-Cog	Mean	0.5	2.5
	Pooled SD	7.4	7.4
CIBIC-Plus	Mean	4.2	4.4
	Pooled SD	1.1	1.1

Subpart (ADAS-Cog) and Clinician’s Interview-Based Impression of Change (CIBIC-Plus).

Suppose that the superiority margins are set at 0 and the noninferiority margins for the ADAS-Cog and CIBIC-Plus endpoints at 0.8 and 0.1, respectively. Based on this information, the superiority and noninferiority test statistics are given by

$$\begin{bmatrix} t_1^{(S)} \\ t_2^{(S)} \end{bmatrix} = \begin{bmatrix} 2.45 \\ 1.65 \end{bmatrix}, \quad \begin{bmatrix} t_1^{(N)} \\ t_2^{(N)} \end{bmatrix} = \begin{bmatrix} 3.43 \\ 2.47 \end{bmatrix}.$$

Further, assuming a one-sided  $\alpha = 0.025$ , the critical values of the Tamhane-Logan procedure are

$$c^{(S)} = t_{0.0125}(326) = 2.25 \text{ and } c^{(N)} = t_{0.025}(326) = 1.97.$$

The rejection region of the superiority-noninferiority procedure is shown in [Figure 4.4](#). The superiority and noninferiority test statistics are in their corresponding rejection regions, i.e.,

$$\max(t_1^{(S)}, t_2^{(S)}) \geq 2.25 \text{ and } \min(t_1^{(N)}, t_2^{(N)}) \geq 1.97,$$

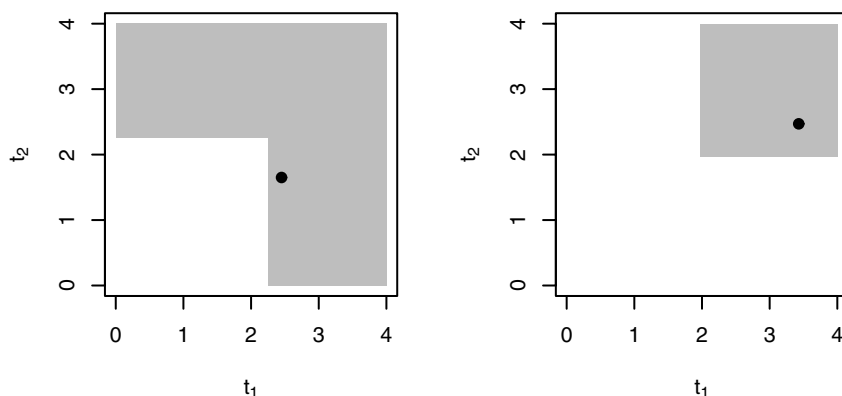
and thus the procedure rejects the global superiority-noninferiority hypothesis. To determine which individual endpoints demonstrate superiority of the treatment,  $t_1^{(S)}$  and  $t_2^{(S)}$  are compared with  $t_{0.025}(326) = 1.97$ . Since only  $t_1^{(S)}$  exceeds this critical constant, superiority is demonstrated only on ADAS-Cog; noninferiority is demonstrated on CIBIC-Plus.

Note that the Tamhane-Logan procedure is monotone in the sense that, if a certain set of test statistics leads to the rejection of the global superiority-noninferiority hypothesis, the hypothesis will be rejected for any set of more extreme test statistics.

It is instructive to compare the Tamhane-Logan superiority-noninferiority procedure to the min test used in the all-or-none testing problem described in Section 4.5. The min test rejects the global superiority hypothesis if

$$\min(t_1^{(S)}, t_2^{(S)}) \geq t_{0.025}(326) = 1.97.$$





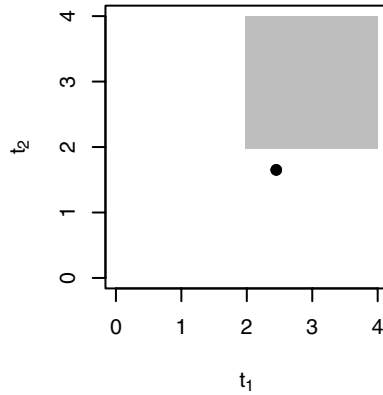
**FIGURE 4.4:** Left panel: Rejection region of the superiority component of the Tamhane-Logan procedure (shaded area) and superiority test statistics (dot). Right panel: Rejection region of the noninferiority component of the Tamhane-Logan procedure (shaded area) and noninferiority test statistics (dot).

The rejection region of the min test is displayed in Figure 4.5. It is clear that the global superiority hypothesis cannot be rejected since the treatment effect for CIBIC-Plus is not significant at the one-sided 0.025 level ( $t_2^{(S)} < 1.97$ ). This serves as an illustration of the fact that the all-or-none testing approach is based on more stringent criteria compared to the superiority-noninferiority testing approach.

## 4.7 Software implementation

This section briefly describes software implementation of the multiple and global procedures discussed in this chapter. The following SAS programs were used in the examples included in this chapter. The programs can be downloaded from the book's Web site (<http://www.multipert.com>).

- Program 2.1 can be used to implement at-least-one procedures for identifying the treatment effect on individual endpoints described in Section 4.3, including the Bonferroni, fixed-sequence and fallback proce-



**FIGURE 4.5:** Rejection region of the min test (shaded area) and superiority test statistics (dot).

dures (note that this program also implements other  $p$ -value-based procedures introduced in Section 2.6). Software implementation of the parametric fallback and 4A procedures is not currently available. Critical values for these procedures are tabulated in Huque and Alosch (2008) and Li and Mehrotra (2008), respectively.

- Program 4.1 implements the OLS and GLS procedures in the osteoarthritis trial example introduced in Section 4.4.2. Program 4.2 performs sample size calculations for the OLS and GLS procedures in this clinical trial example.
- Program 4.3 implements the Tamhane-Logan superiority-noninferiority procedure in the Alzheimer's disease trial example (Section 4.6).

## Acknowledgements

Ajit C. Tamhane's research was supported by grants from the National Heart, Lung and Blood Institute.

# Chapter 5

---

## *Gatekeeping Procedures in Clinical Trials*

Alex Dmitrienko

*Eli Lilly and Company*

Ajit C. Tamhane

*Northwestern University*

---

### 5.1 Introduction

Multiple objectives pursued in clinical trials typically exhibit a hierarchical structure; e.g., they can be divided into primary and secondary objectives (for a detailed classification of primary and secondary endpoints, see D'Agostino, 2000). The primary objective is typically formulated in terms of the primary analysis which describes the most important features of the treatment. In most registration trials, the primary analysis determines the overall outcome of the trial, provides the basis for the regulatory claim and is included in the product label. Secondary analyses (including secondary endpoints and subgroup analyses) play a supportive role and provide additional information for prescribing physicians, patients, payers, etc.

Two approaches to the control of the false positive rate for the family of primary and secondary analyses can be considered in a clinical trial setting:

- Approach 1. The false positive rate is not protected. Regulatory agencies do not always require a strict control of the false positive rate. An example is the case of a small number of secondary analyses or secondary analyses that are highly correlated with the primary analysis.
- Approach 2. The familywise error rate (FWER) associated with the primary and secondary analyses is controlled (FWER is defined in Section 2.2.1).

Although multiplicity adjustments are not mandatory in registration studies to justify the inclusion of secondary endpoints or analyses in the product label, control of the Type I error probability (Approach 2) is becoming increasingly important. This approach is used by regulatory agencies to define the

acceptable statistical risk of false efficacy claims in registration trials. Gatekeeping methods described in this chapter offer a solution to this multiplicity problem. These methods enable the trial's sponsor to

- Control the risk of spurious conclusions (e.g., false efficacy claims) with respect to multiple ordered analyses.
- Take into account the hierarchical structure of the multiple testing problem and examine ordered analyses in a sequential manner beginning with the primary analyses. The gatekeeping methodology is consistent with a regulatory view that findings with respect to secondary/supportive objectives; e.g., secondary endpoints, are meaningful only if the primary objective is met (O'Neill, 1997).

For more information about the use of gatekeeping procedures in a clinical trial setting and literature review in this area of multiple comparison research, see Dmitrienko et al. (2005, Chapter 2) and Dmitrienko and Tamhane (2007).

This chapter begins with motivating examples and a review of gatekeeping procedures in Section 5.2. The next three sections provide a detailed description of three classes of gatekeeping procedures: serial (Section 5.3), parallel (Section 5.4) and tree-structured (Section 5.5). Each section includes a discussion of relevant methodology and clinical trial examples. The last section (Section 5.6) describes available software tools for implementing gatekeeping procedures in clinical trials.

---

## 5.2 Motivating examples

To construct a gatekeeping procedure, one first needs to define two or more families of analyses, for example, a family of primary endpoints and a family of secondary endpoints. Each family (except for the last one) serves as a gatekeeper in the sense that one must pass it to perform analyses in the next family.

In this section we will present clinical trial examples that motivate the use of gatekeeping methods in clinical trials and also set the stage for the review of main classes of gatekeeping procedures in Sections 5.3–5.5.

As a side note, gatekeeping procedures discussed in this section focus on multiplicity adjustments in a single trial. In the context of registration/marketing authorization packages that normally include two confirmatory trials with similar sets of primary and secondary analyses, gatekeeping methods can be applied independently to each trial. This approach will ensure Type I error rate control within each confirmatory trial. To justify the inclusion of secondary findings into the product label, the trial's sponsor can use consistency arguments and demonstrate that multiplicity-adjusted primary

and secondary analyses lead to similar conclusions in both trials. It is worth noting that regulatory guidelines do not currently discuss rules for combining secondary findings across several confirmatory trials.

### 5.2.1 Clinical trials with serial gatekeepers

We will begin with a two-family testing problem arising in clinical trials with noninferiority and superiority objectives (this example is based on Dmitrienko and Tamhane, 2007). Consider a trial in patients with Type II diabetes with three treatment groups, Group A (a new formulation of an insulin therapy), Group B (a standard formulation) and Group A+B (a combination of the formulations). The following two scenarios will be examined:

- Scenario 1. Noninferiority and superiority tests are carried out sequentially for the comparison of A versus B.
- Scenario 2. A noninferiority test for the comparison of A versus B is carried out first followed by a superiority test for the same comparison and a noninferiority test for the comparison of A+B versus B.

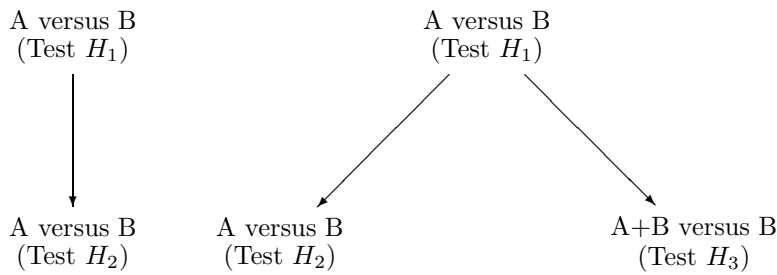
Let  $\delta_1$  and  $\delta_2$  denote the true treatment differences for the comparisons of A versus B and A+B versus B, respectively. The three sets of null and alternative hypotheses arising in this problem are defined as follows:

- A versus B (noninferiority),  $H_1 : \delta_1 \leq -\gamma_1$  versus  $K_1 : \delta_1 > -\gamma_1$ , where  $\gamma_1$  is a positive non-inferiority margin for the comparison of A versus B.
- A versus B (superiority),  $H_2 : \delta_1 \leq 0$  versus  $K_2 : \delta_1 > 0$ .
- A+B versus B (noninferiority),  $H_3 : \delta_2 \leq -\gamma_2$  versus  $K_3 : \delta_2 > -\gamma_2$ , where  $\gamma_2$  is a positive non-inferiority margin for the comparison of A+B versus B.

The testing procedures used in the two scenarios are depicted in [Figure 5.1](#). In both scenarios, testing begins with the first family that includes the test for  $H_1$ . This family serves as a serial gatekeeper in the sense that all hypotheses of no treatment effect must be rejected in the first family to proceed to the second family (in this case there is only one hypothesis in the first family). In Scenario 1, the second family includes the test for  $H_2$ ; in Scenario 2, this family includes the tests for  $H_2$  and  $H_3$ .

It is important to note that, even though two tests are performed in Scenario 1, no multiplicity adjustment is needed. Both tests can be carried out at the same  $\alpha$  level, where  $\alpha$  is the pre-specified FWER, e.g.,  $\alpha = 0.05$ . This is due to the fact that this testing procedure is a special case of the fixed-sequence approach described in Section 2.6.3.

It is instructive to compare the straightforward multiple testing problem in Scenario 1 to the more complex one in Scenario 2. Although the two settings



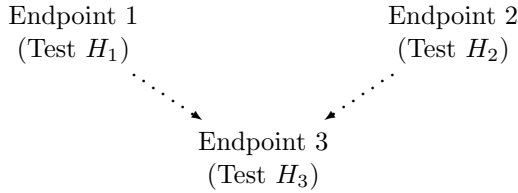
**FIGURE 5.1:** Decision trees in the combination-therapy clinical trial example (Scenario 1, left panel; Scenario 2, right panel).

look quite similar at first glance, one can no longer avoid multiplicity adjustments in Scenario 2. To see this, suppose that no multiplicity adjustment is performed; i.e., all three tests are carried out at the  $\alpha$  level and assume that  $\delta_1 = 0$  (the new formulation is equivalent to the standard formulation) and  $\delta_2 \leq -\gamma_2$  (the combination is inferior to the standard formulation). Under this set of assumptions, if the noninferiority margin for the comparison of A versus B is very wide ( $\gamma_1$  is large), one is virtually assured of observing a significant outcome for the first test and passing the gatekeeper. This means that the original multiple testing problem will simplify to the problem of testing  $H_2$  and  $H_3$ . The two tests are carried at the unadjusted  $\alpha$  level and thus the probability of at least one incorrect conclusion will be inflated. This example illustrates that multiplicity adjustments are needed in general for multi-family testing problems.

The multiple testing problem considered in this section is a two-family problem in which the first family serves as a serial gatekeeper. Serial gatekeepers are often found in clinical trials with multiple ordered endpoints. For example, in a clinical trial with a single primary endpoint and several key secondary endpoints, the endpoints may be arranged in a sequence. In this case, each endpoint defines a family and serves as a serial gatekeeper for the next family in the sequence. Multiple testing procedures that control the FWER in problems with serial gatekeepers are discussed in Section 5.3.

**5.2.2 Clinical trials with parallel gatekeepers**

To introduce parallel gatekeepers, consider an osteoporosis/breast cancer clinical trial in postmenopausal women that investigates the efficacy of a novel treatment compared to a placebo control (this example is based on Cummings et al., 1999; Ettinger et al., 1999). The treatment effect is evaluated using two primary endpoints, incidence of vertebral fractures (Endpoint 1) and incidence of breast cancer (Endpoint 2) and an important secondary endpoint, incidence



**FIGURE 5.2:** Decision tree in the osteoporosis/breast cancer clinical trial example. Dotted lines are used to indicate that only one test in the first family needs to be significant to perform the test in the second family.

of non-vertebral fractures (Endpoint 3). Let  $\delta_i$  denote the true treatment difference for the  $i$ th endpoint. The associated null hypothesis,  $H_i : \delta_i \leq 0$ , is tested against a superiority hypothesis,  $K_i : \delta_i > 0$ ,  $i = 1, 2, 3$ .

The first family includes the primary tests (tests for  $H_1$  and  $H_2$ ) and the second family includes the test for  $H_3$ . Each primary endpoint is associated with an independent regulatory claim and the trial will be declared positive if there is evidence of a beneficial treatment effect for at least one primary endpoint. Using mathematical terminology, the first family serves as a parallel gatekeeper; i.e., at least one hypothesis needs to be rejected in this family to pass the gatekeeper and carry out the test in the second family. The testing procedure is displayed in Figure 5.2.

As in the two-family problem described in Section 5.2.1, it is easy to show that an appropriate multiplicity adjustment strategy is required in this case to preserve the FWER. A naive strategy can be set up as follows:

- Since there are two tests in the first family, a multiple test is used to control the Type I error rate within this family, e.g., Bonferroni test ( $H_1$  and  $H_2$  are tested at the  $\alpha/2$  level).
- If one or more tests in the first family are significant,  $H_3$  is tested at the  $\alpha$  level (after all, there is only one test in the second family).

To verify whether this approach prevents Type I error rate inflation, we can compute the probability of at least one erroneous conclusion when  $\delta_1$  is very large but  $\delta_2 = 0$  and  $\delta_3 = 0$ .

Since the treatment is superior to placebo with a large margin for Endpoint 1, the testing procedure is virtually guaranteed to pass the gatekeeper and the test for  $H_3$  will be carried out almost all of the time. As a result, the three-endpoint problem collapses to a two-endpoint problem in which  $H_2$  is tested at the  $\alpha/2$  level and  $H_3$  is tested at the  $\alpha$  level. It is clear that the probability of at least one incorrect conclusion will be greater than  $\alpha$  (unless the test statistics associated with  $H_2$  and  $H_3$  are perfectly correlated). In other words, even though the naive strategy protects the Type I error rate within each family, the overall Type I error rate ends up being inflated. This clinical trial

example shows that a more sophisticated multiple comparison procedure (that goes beyond Type I error rate control within each family) may be required in trials with hierarchically ordered analyses.

In general, parallel gatekeepers could be utilized in clinical trials with several primary endpoints where each endpoint defines a successful trial outcome; e.g., each endpoint is associated with its own regulatory claim. In addition, parallel gatekeepers could be used in trials with multiple doses of a treatment tested against a control, e.g., a placebo or active control. In this case, the dose-control hypotheses corresponding to higher dose levels could be included in the first family that serves as a parallel gatekeeper for the family containing the other dose-control hypotheses. Section 5.4 introduces a general class of procedures that control the FWER in trials with parallel gatekeepers.

### 5.2.3 Clinical trials with tree-structured gatekeepers

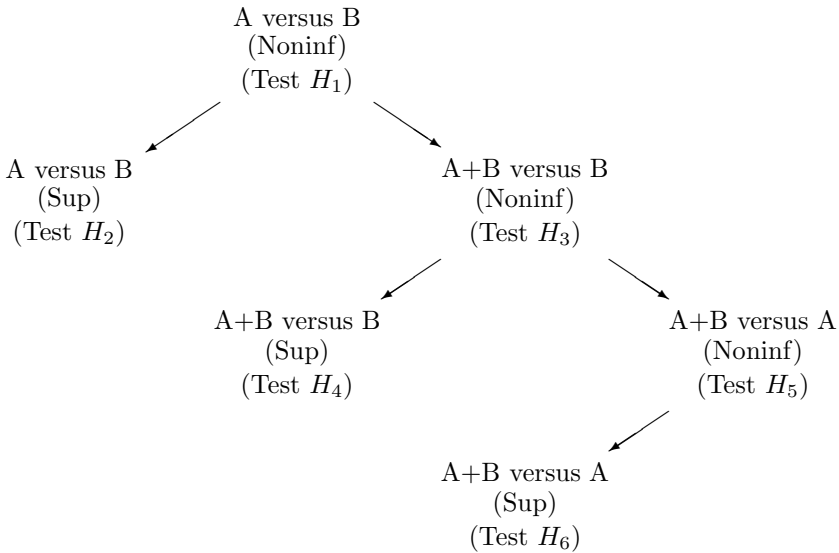
Tree-structured gatekeeping procedures (or, simply, tree gatekeeping procedures) are used in clinical trials with multiple analyses that form a complex hierarchical structure. This includes structures with logical relationships among the analyses that go beyond more basic hierarchical structures associated with serial and parallel gatekeeping methods.

To illustrate, we will consider an extension of the combination-therapy clinical trial example given in Section 5.2.1. In this example, there are six tests that are carried out in four stages as shown below:

- Stage 1. A versus B (test of the noninferiority hypothesis  $H_1$ ).
- Stage 2. A versus B (test of the superiority hypothesis  $H_2$ ) and A+B versus B (test of the noninferiority hypothesis  $H_3$ ).
- Stage 3. A+B versus B (test of the superiority hypothesis  $H_4$ ) and A+B versus A (test of the noninferiority hypothesis  $H_5$ ).
- Stage 4. A+B versus A (test of the superiority hypothesis  $H_6$ ).

A decision tree associated with this testing strategy is displayed in [Figure 5.3](#). The tree exhibits fairly complex logical relationships among the tests. In the parallel gatekeeping example in Section 5.2.2, the secondary test (test for  $H_3$ ) was logically related to both primary tests (tests for  $H_1$  and  $H_2$ ). In this case, each test at Stages 3 and 4 is logically related to only one test carried out at the previous stage. For example,  $H_4$  will be tested if and only if the test for  $H_3$  is significant and the outcome of the test for  $H_2$  is ignored. Gatekeeping procedures for problems with logical restrictions of this kind are known as tree gatekeeping procedures. Using a counterexample similar to the one given in Section 5.2.2, it is easy to show that Type I error rate control within each family does not, in general, guarantee control of the FWER in problems with logical restrictions.





**FIGURE 5.3:** Decision tree in the combination-therapy clinical trial example (Noninf, Noninferiority; Sup, Superiority).

Tree gatekeepers arise in clinical trials with multiple objectives, e.g., multiple endpoints or multiple subgroups, when logical dependencies exist among the null hypotheses associated with these objectives. It is shown in Section 5.5 how to construct tree gatekeeping procedures that take these logical relationships into account and protect the FWER.

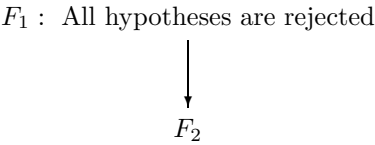
### 5.3 Serial gatekeeping procedures

Sections 5.3–5.5 give a comprehensive review of the three classes of gatekeeping procedures, including underlying theory, clinical trial examples and implementation details.

The following notation will be used in the three sections. Consider a clinical trial with multiple, hierarchically ordered objectives/analyses. To account for the hierarchical ordering, the analyses are grouped into  $m$  families denoted by  $F_1, \dots, F_m$ . Each family includes null hypotheses corresponding to the analyses at the same level in the hierarchy; e.g., the hypotheses in  $F_1$  may be related to a set of primary analyses and the hypotheses in  $F_2$  may represent

**TABLE 5.1:** Families of null hypotheses corresponding to multiple, hierarchically ordered objectives.

Family	Null hypotheses	Hypothesis weights	Raw $p$ -values
$F_1$	$H_{11}, \dots, H_{1n_1}$	$w_{11}, \dots, w_{1n_1}$	$p_{11}, \dots, p_{1n_1}$
$\dots$	$\dots$	$\dots$	$\dots$
$F_i$	$H_{i1}, \dots, H_{in_i}$	$w_{i1}, \dots, w_{in_i}$	$p_{i1}, \dots, p_{in_i}$
$\dots$	$\dots$	$\dots$	$\dots$
$F_m$	$H_{m1}, \dots, H_{mn_m}$	$w_{m1}, \dots, w_{mn_m}$	$p_{m1}, \dots, p_{mn_m}$



**FIGURE 5.4:** A problem with a serial gatekeeper ( $F_1$  is a serial gatekeeper for  $F_2$ ).

secondary analyses. As was stated in Section 5.2, each family (except for the last one) serves as a gatekeeper for the families placed later in the sequence.

The hypotheses included in the  $m$  families are shown in Table 5.1. Further,  $w_{i1}, \dots, w_{in_i}$  are the weights representing the importance of hypotheses within  $F_i$  (the weights are non-negative and  $w_{i1} + \dots + w_{in_i} = 1$ ) and  $p_{i1}, \dots, p_{in_i}$  are the associated raw  $p$ -values. Multiplicity adjusted  $p$ -values for the hypotheses in  $F_i$  are denoted by  $\tilde{p}_{i1}, \dots, \tilde{p}_{in_i}$  (note that the adjusted  $p$ -values are defined with respect to all  $m$  families rather than any individual family).

5.3.1 General serial gatekeeping framework

A family is termed a *serial gatekeeper* if all hypotheses must be rejected within that family in order to proceed to the next family in the sequence (see Figure 5.4). In other words, if  $F_i$ ,  $i = 1, \dots, m - 1$ , is a serial gatekeeper, hypotheses in  $F_{i+1}$  are tested if and only if

$$\max_{j=1, \dots, n_i} \tilde{p}_{ij} \leq \alpha.$$

A clinical trial example with a serial gatekeeper was given in Section 5.2.1.

Serial gatekeeping procedures were studied by Maurer, Hothorn and Lehmacher (1995), Bauer et al. (1998) and Westfall and Krishen (2001). Most commonly, serial gatekeepers are encountered in trials where endpoints can be ordered from most important to least important:

- The adalimumab trial in patients with rheumatoid arthritis (Keystone et al., 2004) tested the effect of adalimumab on three endpoints that

were ordered and examined sequentially: symptomatic response, disease progression and physical function.

- Hierarchical arrangements of endpoints are often used in oncology trials, e.g., overall survival duration, progression-free survival duration, tumor response rate, time to treatment failure and duration of tumor response.

Serial gatekeeping procedures are widely used in clinical trials, mainly due to the fact that they do not require an adjustment for multiplicity. Note that serial gatekeeping procedures are closely related to the fixed-sequence test introduced in Section 2.6.3 (in fact, these procedures simplify to the fixed-sequence test if each family includes a single hypothesis). This approach to testing ordered endpoints is described in the CPMP guidance document on multiplicity issues in clinical trials (CPMP, 2002).

### 5.3.2 Serial gatekeeping procedures with a single decision-making branch

In their most basic form, serial gatekeeping procedures can be applied to problems in which multiple analyses define a single sequence of hypotheses. We refer to these serial gatekeeping procedures as single-branch procedures.

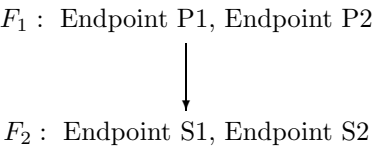
A single-branch procedure for multiple families of analysis is defined as follows. Within each family  $F_i$ ,  $i = 1, \dots, m-1$ , hypotheses are tested at the nominal  $\alpha$  level. For example, the hypotheses in  $F_i$  can be tested using an intersection-union (IU) test (Section 2.3.2); i.e., all hypotheses are rejected in  $F_i$  if  $p_{ij} \leq \alpha$ ,  $j = 1, \dots, n_i$ , and all hypotheses are retained otherwise. Any FWER-controlling test can be used in  $F_m$ , including all popular multiple tests described in Sections 2.6–2.8.

Multiplicity adjustments are commonly summarized using adjusted  $p$ -values for hypotheses of interest. Adjusted  $p$ -values for single-branch procedures are easy to compute using the Westfall-Young definition discussed in Section 2.4.1. Assume that the IU test is used in  $F_1, \dots, F_{m-1}$ . Let  $p_i^*$  denote the largest  $p$ -value in  $F_i$ ,  $i = 1, \dots, m-1$ , and  $p'_{mj}$  denote the adjusted  $p$ -value for  $H_{mj}$  produced by the test used in the last family,  $j = 1, \dots, n_m$ . The adjusted  $p$ -value for  $H_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ , is given by:

$$\tilde{p}_{ij} = \begin{cases} \max(p_1^*, \dots, p_i^*) & \text{if } i = 1, \dots, m-1, \\ \max(p'_{mj}, p_1^*, \dots, p_{i-1}^*) & \text{if } i = m. \end{cases}$$

### Alzheimer's disease clinical trial example

The Alzheimer's disease clinical trial example from Dmitrienko and Tamhane (2007) serves as an example of a single-branch problem with a serial gatekeeper. In this example, the efficacy profile of an experimental treatment is compared to that of a placebo using four endpoints:



**FIGURE 5.5:** Single-branch serial gatekeeping procedure in the Alzheimer’s disease clinical trial example.

**TABLE 5.2:** Serial gatekeeping procedure in the Alzheimer’s disease clinical trial example. The asterisk identifies the adjusted  $p$ -values that are significant at the two-sided 0.05 level.

Family	Endpoint	Raw $p$ -value	Adjusted $p$ -value
$F_1$	P1	0.023	0.023*
$F_1$	P2	0.018	0.023*
$F_2$	S1	0.014	0.028*
$F_2$	S2	0.106	0.106

- The primary endpoints include a cognitive impairment endpoint, ADAS-Cog (Endpoint P1), and a clinical global performance endpoint, CIBIC (Endpoint P2).
- Two secondary endpoints are also examined in this trial, a biochemical endpoint (Endpoint S1) and an imaging endpoint (Endpoint S2).

The hypotheses for the primary and secondary endpoints are included in  $F_1$  and  $F_2$ , respectively. Since a trial for the treatment of Alzheimer’s disease is normally declared successful only if both primary endpoints are significant (Reisberg et al., 2003; Reines et al., 2004),  $F_1$  serves as a serial gatekeeper (see Figure 5.5).

To illustrate the implementation of the serial gatekeeping procedure, Table 5.2 displays the two-sided raw  $p$ -values produced by the four tests in this clinical trial example as well as adjusted  $p$ -values. The hypotheses in  $F_1$  are tested using the IU test and both of them are rejected at the 0.05 level. Because of this, the procedure can pass the gatekeeper and test the hypotheses in  $F_2$ . The Holm test is carried out in  $F_2$  and the adjusted  $p$ -values for Endpoints S1 and S2 are given by 0.028 and 0.106, respectively. Endpoint S1 is significant at the 0.05 level, whereas Endpoint S2 is not. Since the serial gatekeeping procedure controls the FWER, the trial’s sponsor can use these results to justify the inclusion of the two primary endpoints as well as one secondary endpoint (Endpoint S1) in the product label.

Serial gatekeeping procedures have a simple structure and are quite appealing in clinical trial applications. However, it is important to bear in mind that these procedures are based on the fixed-sequence approach and thus they

need to be considered only if there is sufficient clinical justification that can be used to prioritize the objectives of interest (pros and cons of the fixed-sequence approach are discussed in Section 4.3.1).

### 5.3.3 Serial gatekeeping procedures with multiple decision-making branches

In the previous section we considered a class of basic single-branch procedures. More complicated examples of serial gatekeeping procedures arise in clinical trials with multiple sequences of hypotheses or multiple decision-making branches, e.g., dose-finding studies with ordered endpoints. In this case, at each fixed dose level, dose-control comparisons for multiple endpoints form a branch within which hypotheses are tested sequentially.

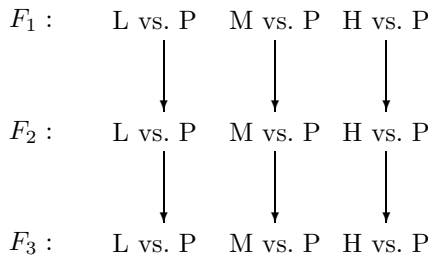
Serial gatekeeping procedures with multiple branches can be constructed based on several multiple tests. Here we will focus on Bonferroni-based procedures (serial gatekeeping procedures based on other tests are briefly discussed in Section 5.3.4). Consider a multiple testing problem with  $m$  families and assume that each one contains  $n$  hypotheses, i.e.,  $n_1 = \dots = n_m = n$ . In this case there are  $n$  branches and the  $j$ th branch includes the hypotheses  $H_{1j}, \dots, H_{mj}$ . Hypotheses within each branch are tested sequentially as follows:

- Consider the  $j$ th branch,  $j = 1, \dots, n$ . The hypothesis  $H_{1j}$  is tested first at an  $\alpha/n$  level. If  $H_{1j}$  is rejected, the next hypothesis in the sequence, i.e.,  $H_{2j}$ , is tested, otherwise testing within this branch stops.
- In general, the hypothesis  $H_{ij}$  is rejected if  $p_{kj} \leq \alpha/n$  for all  $k = 1, \dots, i$ .

FWER control for the Bonferroni-based procedure is discussed in Quan, Luo and Capizzi (2005). Adjusted  $p$ -values for serial gatekeeping procedures with multiple branches can be found using the direct calculation algorithm defined in Section 5.4.2.

### Type II diabetes clinical trial example

A multiple testing problem with three branches was described in Dmitrienko et al. (2006a) and Dmitrienko et al. (2007). The Type II diabetes clinical trial considered in these papers is conducted to compare three doses of an experimental treatment (labeled L, M and H) versus placebo (labeled P). Each dose-placebo test is carried out with respect to three ordered endpoints: hemoglobin A1c (Endpoint E1), fasting serum glucose (Endpoint E2) and HDL cholesterol (Endpoint E3). The E2 tests are restricted to the doses at which Endpoint E1 is significant and, similarly, the E3 tests are carried out only for the doses at which the E1 and E2 tests are both significant. Logical restrictions of this kind facilitate drug labeling and, in addition, improve the power of clinically relevant secondary dose-placebo tests. The resulting



**FIGURE 5.6:** Three-branch serial gatekeeping procedure with three families of hypotheses in the Type II diabetes clinical trial example ( $F_1$ , Endpoint E1;  $F_2$ , Endpoint E2;  $F_3$ , Endpoint E3).

decision tree has three branches (see Figure 5.6) and the fixed-sequence approach is applied within each branch. The branches are “connected” using the Bonferroni test as described below.

To define the three-branch procedure, the hypotheses  $H_{i1}$  (H-P comparison),  $H_{i2}$  (M-P comparison) and  $H_{i3}$  (L-P comparison) for the  $i$ th endpoint are included in  $F_i$ ,  $i = 1, 2, 3$ . The hypotheses are equally weighted within each family and the FWER is set at a two-sided  $\alpha = 0.05$ . The hypotheses within the three branches are tested sequentially using the Bonferroni-based procedure.

The two-sided raw and adjusted  $p$ -values in this clinical trial example are summarized in Table 5.3. The adjusted  $p$ -values are computed using the direct-calculation algorithm with  $K = 100,000$ . Note that only Doses M and H are significantly different from placebo for the primary endpoint (Endpoint E1) and thus the remaining branch corresponding to the L-P comparison is eliminated at the first stage of the procedure. At the second stage, the dose-placebo comparisons for Endpoint E2 are performed only for the dose levels at which Endpoint E1 is significant, i.e., Doses M and H. There is no evidence of a significant effect at Dose M compared to placebo for Endpoint E2 and thus testing within that branch stops. At the last stage, Dose H is tested against placebo for Endpoint E3. This test is significant and thus we conclude that Dose H is superior to placebo for all three endpoints whereas Dose M is superior to placebo only for Endpoint E1.

**5.3.4 Other serial gatekeeping procedures**

In general, sponsors of clinical trials may consider more complicated serial gatekeeping procedures, including multiple-branch with unequal length branches (this setting is encountered in trials that compare a treatment to multiple controls). Further, multiple-branch serial gatekeeping procedures can be constructed based on other multiple tests, e.g., the Hochberg test (Quan,

**TABLE 5.3:** Serial gatekeeping procedure in the Type II diabetes clinical trial example. The asterisk identifies the adjusted  $p$ -values that are significant at the two-sided 0.05 level.

Family	Endpoint	Comparison	Raw $p$ -value	Adjusted $p$ -value
$F_1$	E1	L vs. P	0.0176	0.0528
$F_1$	E1	M vs. P	0.0108	0.0324*
$F_1$	E1	H vs. P	0.0052	0.0156*
$F_2$	E2	L vs. P	0.0128	0.0528
$F_2$	E2	M vs. P	0.0259	0.0777
$F_2$	E2	H vs. P	0.0093	0.0279*
$F_3$	E3	L vs. P	0.0511	0.1533
$F_3$	E3	M vs. P	0.0058	0.0777
$F_3$	E3	H vs. P	0.0099	0.0297*

Luo and Capizzi, 2005) or Dunnett test (Dmitrienko et al., 2006a; Dmitrienko, Tamhane and Liu, 2008).

### 5.4 Parallel gatekeeping procedures

This section gives an overview of multiplicity adjustment methods used in parallel gatekeeping procedures.

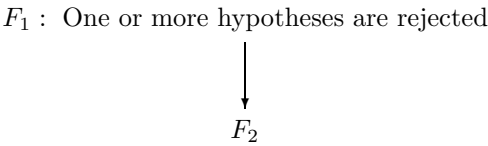
#### 5.4.1 General parallel gatekeeping framework

Family  $F_i$  is termed a *parallel gatekeeper* if at least one significant result must be observed in this family, i.e., one or more hypotheses must be rejected in  $\{H_{i1}, \dots, H_{in_i}\}$ , to proceed to  $F_{i+1}$ ,  $i = 1, \dots, m - 1$  (see Figure 5.7). In other words, if testing is performed at the  $\alpha$  level, the gatekeeper is passed if and only if

$$\min_{j=1, \dots, n_i} \tilde{p}_{ij} \leq \alpha.$$

As an illustration, a multiple testing problem with a parallel gatekeeper was discussed in Section 5.2.2. Other examples can be found in clinical trials with multiple primary endpoints when each endpoint provides independent proof of efficacy and can lead to a regulatory claim, e.g., the acute respiratory distress syndrome clinical trial (Dmitrienko, Offen and Westfall, 2003, Section 4) with two primary endpoints, number of ventilator-free days and 28-day all-cause mortality, or the EPHESUS trial (Pitt et al., 2003) that utilized two primary endpoints, all-cause mortality and cardiovascular mortality plus cardiovascular hospitalization.

The parallel gatekeeping methods were introduced in Dmitrienko, Offen



**FIGURE 5.7:** A problem with a parallel gatekeeper ( $F_1$  is a parallel gatekeeper for  $F_2$ ).

and Westfall (2003) who considered a Bonferroni-based procedure derived using the closure principle (see Section 2.3.3). Since this method relies on a complete enumeration of all intersection hypotheses in the closed family associated with  $F_1, \dots, F_m$ , the resulting parallel gatekeeping procedures may lack transparency and their implementation can be computationally intensive since it takes order- $2^n$  steps to test  $n$  hypotheses.

Further research in this area revealed that a broad class of parallel gatekeeping procedures have a stepwise form (Dmitrienko et al., 2006b; Hommel, Bretz and Maurer, 2007; Guilbaud, 2007; Dmitrienko, Tamhane and Wiens, 2008). This property streamlines their implementation and interpretation by clinical trial practitioners (US Food and Drug Administration statisticians have repeatedly emphasized the importance of multiple testing procedures that can be understood by clinicians). In this section we will focus on multistage parallel gatekeeping procedures developed in Dmitrienko, Tamhane and Wiens (2008).

**5.4.2 Multistage parallel gatekeeping procedures**

We will begin by introducing two concepts that play a key role in the framework for constructing multistage parallel gatekeeping procedures: the *error rate function* of a multiple test and *separable* multiple tests.

Consider the problem of testing a single family of  $n$  null hypotheses  $H_1, \dots, H_n$ . For any subset  $I$  of the index set  $N = \{1, 2, \dots, n\}$ , the error rate function  $e(I)$  of a multiple test is the maximum probability of making at least one Type I error when testing the hypotheses  $H_i, i \in I$ , i.e.,

$$e(I) = \sup_{H_I} P \left\{ \bigcup_{i \in I} (\text{Reject } H_i) \middle| H_I \right\}.$$

Here the supremum of the probability is computed over the entire parameter space corresponding to the null hypothesis

$$H_I = \bigcap_{i \in I} H_i.$$

An exact expression for  $e(I)$  is generally difficult to derive and an easily



computable upper bound on  $e(I)$  can be used instead. For example, the upper bound for the error rate function of the Bonferroni test is given by  $\alpha|I|/n$ , where  $|I|$  is the number of elements in the index set  $I$ . To simplify notation in this section, if an exact expression for  $e(I)$  is available, we will use the original error rate function; otherwise  $e(I)$  will denote an upper bound on the error rate function.

Note that, by definition,  $e(\emptyset) = 0$  and  $e(N) = \alpha$ , where  $\alpha$  is the FWER. In addition, it is natural to require that the error rate function be monotone, i.e.,  $e(I) \leq e(J)$  if  $I \subseteq J$ . If the monotonicity condition is not satisfied, one can easily enforce monotonicity by using the following upper bound in place of the original error rate function

$$e^*(I) = \max_{I' \subseteq I} e(I').$$

It is easy to see that  $e^*(I)$  is a monotone error rate function.

A multiple test meets the *separability condition* (and is termed *separable*) if its error rate function is strictly less than (separates from)  $\alpha$  unless all hypotheses are true, i.e.,

$$e(I) < \alpha \text{ for all } I \subset N.$$

The Bonferroni test clearly satisfies this condition since  $e(I) < \alpha$  for any index set  $I$  with less than  $n$  elements.

## Truncated multiple tests

It is easy to show that most popular multiple tests, with the exception of the Bonferroni test, do not meet the separability condition. To construct separable multiple tests, Dmitrienko, Tamhane and Wiens (2008) proposed *truncated* versions of popular tests by taking a convex combination of their critical values with the critical values of the Bonferroni test. As a result, a truncated test is uniformly more powerful than the Bonferroni test but uniformly less powerful than the original test. As an illustration, we will define the truncated Holm and Hochberg tests in this section. Truncated versions of other tests; e.g., the fallback and Dunnett tests, and their error rate functions are given in Dmitrienko, Tamhane and Wiens (2008).

To define the two truncated tests, consider the ordered  $p$ -values,  $p_{(1)} \leq \dots \leq p_{(n)}$  and let  $H_{(1)}, \dots, H_{(n)}$  denote the corresponding hypotheses. The truncated Holm test is a step-down test based on the following critical values:

$$c_i = \left[ \frac{\gamma}{n-i+1} + \frac{1-\gamma}{n} \right] \alpha, \quad i = 1, \dots, n,$$

where  $0 \leq \gamma < 1$  is the truncation fraction. In other words, the truncated Holm test begins by testing  $H_{(1)}$  at a  $c_1$  level. If  $p_{(1)} \leq c_1$ , this hypothesis is rejected and the next hypothesis is examined. In general, the truncated test

rejects  $H_{(i)}$  if  $p_{(j)} \leq c_j$  for all  $j \leq i$  and retains  $H_{(i)}, \dots, H_{(n)}$  otherwise. This test simplifies to the Bonferroni test if  $\gamma = 0$  and to the regular Holm test if  $\gamma = 1$ . The error rate function of the truncated Holm test is given by  $e(I) = [\gamma + (1 - \gamma)|I|/n]\alpha$  if  $|I| > 0$  and 0 otherwise.

The truncated Hochberg test utilizes the same set of critical values but it is set up as a step-up test. For  $\gamma = 0$  and  $\gamma = 1$ , this truncated test reduces to the Bonferroni and regular Hochberg tests, respectively. The error rate function of the truncated Hochberg test is given by

$$e(I) = 1 - P \left\{ p_{(i)}(I) > \left[ \frac{\gamma}{|I| - i + 1} + \frac{1 - \gamma}{n} \right] \alpha \text{ for all } i \in I \right\}$$

if  $|I| > 0$  and 0 if  $|I| = 0$ . Here  $p_{(i)}(I)$  denotes the  $i$ th ordered  $p$ -value associated with the index set  $I$ ,  $i = 1, \dots, |I|$ . The calculation of this error rate function is discussed in the Appendix. In the case of two hypotheses, the error rate function of the truncated Hochberg test is equivalent to that of the truncated Holm test, i.e.,  $e(I) = [\gamma + (1 - \gamma)|I|/2]\alpha$  if  $|I| > 0$  and 0 otherwise.

## Multistage testing algorithm

Consider again families  $F_1, \dots, F_m$  corresponding to multiple analyses in a clinical trial and assume that  $F_i$ ,  $i = 1, \dots, m - 1$ , is a parallel gatekeeper (as a side note, this framework also includes serial gatekeepers since any serial gatekeeper can be expressed as a series of single-hypothesis families). Let  $A_i$  denote the index set corresponding to the retained hypotheses in  $F_i$  and  $e_i(I)$  denote the error rate function for the test used in  $F_i$ ,  $i = 1, \dots, m - 1$ . The following algorithm defines a broad class of parallel gatekeeping procedures with a stepwise structure.

- Family  $F_1$ . The hypotheses are tested at an  $\alpha_1$  level using any FWER-controlling separable multiple test, where  $\alpha_1 = \alpha$ .
- Family  $F_i$ ,  $i = 2, \dots, m - 1$ . The hypotheses are tested at an  $\alpha_i$  level using any FWER-controlling separable multiple test, where

$$\alpha_i = \alpha_{i-1} - e_{i-1}(A_{i-1}).$$

- Family  $F_m$ . The hypotheses are tested at an  $\alpha_m$  level using any FWER-controlling multiple test that controls the FWER within  $F_m$ , where

$$\alpha_m = \alpha_{m-1} - e_{m-1}(A_{m-1}).$$

Gatekeeping procedures constructed using this algorithm satisfy two important conditions:

- Parallel gatekeeping condition: A null hypothesis in  $F_i$ ,  $i = 2, \dots, m$ , cannot be rejected if all hypotheses in  $F_{i-1}$  are retained. This is a direct consequence of the fact that the “unused” Type I error rate,  $\alpha_i = \alpha_{i-1} - e_{i-1}(A_{i-1}) = 0$  if  $A_{i-1} = N_{i-1}$  (all hypotheses in  $F_{i-1}$  are retained).

- Independence condition: A decision to reject a null hypothesis in  $F_i$ ,  $i = 1, \dots, m-1$ , is independent of decisions made in  $F_{i+1}, \dots, F_m$  due to the stepwise form of gatekeeping procedures. This condition is consistent with the regulatory requirement that the primary analyses in a registration clinical trial be independent of secondary analyses. However, if the independence condition is not considered critical, one can construct gatekeeping procedures that have more power for tests in the first family. For a discussion of the independence condition, see Dmitrienko et al. (2005, Sections 2.7.2–2.7.3) and Hommel, Bretz and Maurer (2007, Section 4).

It follows from the multistage testing algorithm that the penalty paid for performing multiple inferences in  $F_i$ ,  $i = 2, \dots, m$ , depends on the number of the hypotheses rejected at earlier stages. Note that  $\alpha_1, \dots, \alpha_m$  is a non-increasing sequence, which implies that one faces higher hurdles later in the sequence unless all hypotheses are rejected in previously examined families. The rate at which  $\alpha_i$  decreases depends on the tests used at each stage of the procedure. As an illustration, assume that the hypotheses in  $F_i$ ,  $i = 1, \dots, m-1$ , are tested using the Bonferroni test. In this case,

$$\alpha_i = \frac{r_{i-1}\alpha_{i-1}}{n_{i-1}}, \quad i = 2, \dots, m,$$

where  $r_{i-1}$  is the number of hypotheses rejected in  $F_{i-1}$ . In other words, the fraction of the FWER used in  $F_i$  is the product of the proportions of rejected hypotheses in  $F_1$  through  $F_{i-1}$ . If the truncated Holm test is used in  $F_i$ ,  $i = 1, \dots, m-1$ ,

$$\alpha_i = \begin{cases} (1 - \gamma_{i-1})r_{i-1}\alpha_{i-1}/n_{i-1} & \text{if } r_{i-1} < n_{i-1}, \\ \alpha_{i-1} & \text{if } r_{i-1} = n_{i-1}, \end{cases} \quad i = 2, \dots, m,$$

where  $\gamma_{i-1}$  is the truncation fraction used in  $F_{i-1}$ . It follows from this formula that  $\gamma_{i-1}$  determines the fraction of  $\alpha_{i-1}$  carried forward (unless all hypotheses are rejected in  $F_{i-1}$  in which case all of  $\alpha_{i-1}$  is carried over to  $F_i$ ). If the truncation fraction is close to 1 and some hypotheses are retained in  $F_{i-1}$ , an extremely small fraction of  $\alpha_{i-1}$  will be carried over to  $F_i$ .

## Computation of adjusted $p$ -values

The Westfall-Young definition of an adjusted  $p$ -value given in Section 2.4.1 can be applied to calculate adjusted  $p$ -values for multistage gatekeeping procedures using the following direct calculation algorithm. This algorithm loops through a grid of significance levels between 0 and 1 to find the lowest level at which each hypothesis is rejected. The adjusted  $p$ -value for  $H_{ij}$  corresponds to the smallest  $k$ ,  $1 \leq k \leq K$ , for which  $H_{ij}$  is rejected at the overall level  $k\alpha/K$ . The algorithm is quite fast since it takes only  $K = 100,000$  iterations to compute adjusted  $p$ -values with four accurate decimal places. In addition,

multistage gatekeeping procedures have a stepwise form and thus each iteration requires order- $n$  operations to test  $n$  hypotheses.

In special cases, a recursive approach can be applied to calculate adjusted  $p$ -values for multistage parallel gatekeeping procedures. For example, Guibaud (2007) obtained a recursive formula for Bonferroni-based multistage parallel gatekeeping procedures. Consider, for simplicity, a multiple testing problem with two families,  $F_1$  and  $F_2$ , and assume that the hypotheses are equally weighted within each family. The hypotheses in  $F_1$  and  $F_2$  are tested using the Bonferroni test and an arbitrary FWER-controlling test, respectively. Let  $p_{1j}$ ,  $j = 1, \dots, n_1$ , denote the raw  $p$ -values for the hypotheses in  $F_1$ . Further, let  $p'_{2j}$ ,  $j = 1, \dots, n_2$ , denote the adjusted  $p$ -values for the hypotheses in  $F_2$  produced by the test used at the second stage of the procedure. The adjusted  $p$ -values in  $F_1$  are given by

$$\tilde{p}_{1j} = \min(1, n_1 p_{1j}), \quad j = 1, \dots, n_1.$$

Now, consider the ordered adjusted  $p$ -values in  $F_1$ , i.e.,  $\tilde{p}_{1(1)} < \dots < \tilde{p}_{1(n_1)}$ . The adjusted  $p$ -values in  $F_2$  are given by

$$\tilde{p}_{2j} = \min_{k=1, \dots, n_1} \max(\tilde{p}_{1(k)}, n_1 p'_{2j}/k), \quad j = 1, \dots, n_2.$$

## General parallel gatekeeping procedures

It is worth emphasizing that different multiple tests, can be used at different stages of the algorithm introduced earlier in this section. This includes truncated versions of all popular multiple tests introduced in Sections 2.6–2.8; e.g.,  $p$ -value-based tests and tests that account for the correlation among the test statistics within each family (parametric and resampling multiple tests).

In addition, this parallel gatekeeping framework can be extended to procedures that account for the correlation across the families. Note that these gatekeeping procedures are constructed using the general closure method and may not have a stepwise form. For example, a closure-based parametric gatekeeping procedure derived from the Dunnett test was developed in Dmitrienko et al. (2006a). This procedure can be employed when the test statistics follow a multivariate normal distribution. Examples include dose-finding clinical trials with multiple normally distributed outcome variables, e.g., the Type II diabetes clinical trial example used in Section 5.3.3. Further, a Bonferroni-based resampling gatekeeping procedure was proposed in Dmitrienko, Offen and Westfall (2003). Unlike parametric procedures, this procedure does not make the normality assumption and can be applied to a broader class of multiple testing problems with a hierarchical structure.

### 5.4.3 Cardiovascular clinical trial example

The multistage parallel gatekeeping framework will be illustrated using a clinical trial example based on the EPHESUS trial (Pitt et al., 2003). This trial

**TABLE 5.4:** Two-sided  $p$ -values in the cardiovascular clinical trial example.

Family	Hypothesis	Endpoint	Raw $p$ -value	
			Scenario 1	Scenario 2
$F_1$	$H_{11}$	P1	0.0121	0.0121
$F_1$	$H_{12}$	P2	0.0337	0.0872
$F_2$	$H_{21}$	S1	0.0084	0.0084
$F_2$	$H_{22}$	S2	0.0160	0.0160

was conducted to assess the effects of eplerenone on morbidity and mortality in patients with severe heart failure. In this clinical trial example, we will consider two families of endpoints:

- Two primary endpoints: all-cause mortality (Endpoint P1) and cardiovascular mortality plus cardiovascular hospitalization (Endpoint P2).
- Two major secondary endpoints: cardiovascular mortality (Endpoint S1) and all-cause mortality plus all-cause hospitalization (Endpoint S2).

The family of primary endpoints serves as a parallel gatekeeper for the family of secondary endpoints. The hypotheses of no treatment effect are defined as follows: The hypotheses  $H_{11}$  (Endpoint P1) and  $H_{12}$  (Endpoint P2) are included in  $F_1$  and the hypotheses  $H_{21}$  (Endpoint S1) and  $H_{22}$  (Endpoint S2) are included in  $F_2$ . The hypotheses are equally weighted within each family and the pre-specified FWER is  $\alpha = 0.05$ . Table 5.4 displays two sets of two-sided  $p$ -values for the four endpoints that will be used in this example (note that these  $p$ -values are used here for illustration only). Under Scenario 1, the effect size is large for both primary endpoints and, under Scenario 2, there is evidence of treatment effect for only one primary endpoint (Endpoint P1).

A two-stage parallel gatekeeping procedure will be set up as follows. The hypotheses in  $F_1$  and  $F_2$  will be tested using the truncated and regular Holm tests, respectively. The truncated Holm test is carried out using four values of the truncation parameter ( $\gamma = 0, 0.25, 0.5$  and  $0.75$ ) to evaluate the impact of this parameter on the outcomes of the four analyses.

To illustrate the process of applying the two-stage gatekeeping procedure, consider Scenario 1 and let  $\gamma = 0.25$ . The hypotheses  $H_{11}$  and  $H_{12}$  are tested using the truncated Holm test at  $\alpha_1 = \alpha = 0.05$ . The smaller  $p$ -value,  $p_{11} = 0.0121$ , is less than

$$[\gamma/2 + (1 - \gamma)/2]\alpha = \alpha/2 = 0.025$$

and thus  $H_{11}$  is rejected. Further, the larger  $p$ -value,  $p_{12} = 0.0337$ , is compared to

$$[\gamma + (1 - \gamma)/2]\alpha = 5\alpha/8 = 0.03125.$$

**TABLE 5.5:** Parallel gatekeeping procedure in the cardiovascular clinical trial example. The tests in  $F_1$  are carried out using the truncated Holm test with  $\gamma = 0, 0.25, 0.5$  and  $0.75$  and the tests in  $F_2$  are carried out using the regular Holm test. The asterisk identifies the adjusted  $p$ -values that are significant at the two-sided  $0.05$  level.

Family	Endpoint	Adjusted $p$ -value			
		$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$	$\gamma = 0.75$
Scenario 1					
$F_1$	P1	0.0242*	0.0242*	0.0242*	0.0242*
$F_1$	P2	0.0674	0.0539	0.0449*	0.0385*
$F_2$	S1	0.0336*	0.0448*	0.0449*	0.0385*
$F_2$	S2	0.0336*	0.0448*	0.0449*	0.0385*
Scenario 2					
$F_1$	P1	0.0242*	0.0242*	0.0242*	0.0242*
$F_1$	P2	0.1744	0.1395	0.1163	0.0997
$F_2$	S1	0.0336*	0.0448*	0.0672	0.0997
$F_2$	S2	0.0336*	0.0448*	0.0672	0.0997

The corresponding hypothesis cannot be rejected since  $p_{12} > 0.03125$ . To find the fraction of  $\alpha$  that can be carried over to the hypotheses in  $F_2$ , note that the set of retained hypotheses in  $F_1$  includes only one hypothesis. Thus,  $|A_1| = 1$ ,  $n = 2$  and

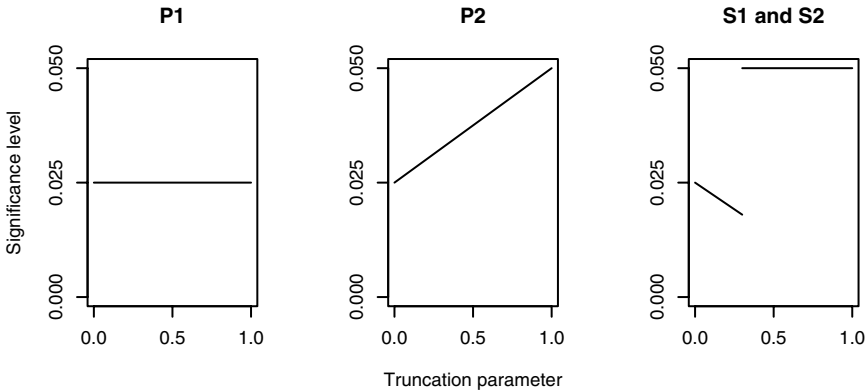
$$\alpha_2 = \alpha_1 - e_1(A_1) = \alpha - [\gamma + (1 - \gamma)|A_1|/n]\alpha = 3\alpha/8 = 0.01875.$$

Applying the regular Holm test in  $F_2$  at  $\alpha_2 = 0.01875$ , it is easy to verify that  $p_{21} < \alpha_2/2$  and  $p_{22} < \alpha_2$ . This implies that the hypotheses  $H_{21}$  and  $H_{22}$  are rejected.

The adjusted  $p$ -values produced by the two-stage gatekeeping procedure are shown in Table 5.5. The adjusted  $p$ -values are computed using the direct-calculation algorithm with  $K = 100,000$ .

As we emphasized earlier in this section, the choice of the truncation parameter  $\gamma$  has a substantial impact on the outcomes of individual tests. It is clear from Table 5.5 that the adjusted  $p$ -values in the primary family ( $F_1$ ) are non-increasing functions of  $\gamma$  (note that the adjusted  $p$ -value for Endpoint P1 is constant because the critical value of the truncated Holm test for the smallest  $p$ -value in  $F_1$  does not actually depend on  $\gamma$ ). However, the relationship between  $\gamma$  and the adjusted  $p$ -values for the secondary endpoints is more complicated. As  $\gamma$  increases, the fraction of  $\alpha$  carried over to the secondary analyses may increase or decrease depending on effect sizes for false hypotheses and this directly influences the adjusted  $p$ -values in  $F_2$ .

In Scenario 1 a small increase in  $\gamma$  from 0 causes an increase in the adjusted  $p$ -values for Endpoints S1 and S2 (compare the columns for  $\gamma = 0$  and  $\gamma = 0.25$ ). Further, when  $\gamma = 0.5$ , these adjusted  $p$ -values stay at the same level and, when  $\gamma = 0.75$ , they drop to 0.037. This is due to the fact



**FIGURE 5.8:** Significance level for Endpoint P1 (left panel), significance level for Endpoint P2 (middle panel) and overall significance level for Endpoints S1 and S2 (right panel) in Scenario 1 as a function of the truncation parameter  $\gamma$ .

that both primary tests achieve significance when  $\gamma \geq 0.5$  and thus the entire  $\alpha$  is propagated to the family of secondary endpoints. To illustrate this phenomenon, Figure 5.8 displays the individual significance levels for the two primary endpoints and overall  $\alpha$  level for the secondary endpoints in Scenario 1 as a function of the truncation parameter.

Figure 5.8 shows that, as was pointed out above, the significance level for Endpoint P1 is constant, the significance level for Endpoint P2 increases with  $\gamma$ , and the overall significance level for Endpoints S1 and S2 is a non-linear function of  $\gamma$  with a jump discontinuity at  $\gamma = 0.3$ . This discontinuity corresponds to the point when the  $p$ -value for Endpoint P2 becomes significant and thus, by the definition of the error rate function of the truncated Holm test, the  $\alpha$  level for the secondary endpoints is set at 0.05. Table 5.5 and Figure 5.8 indicate that, when the effect sizes of both primary endpoints are large, the overall power is maximized by selecting a value of  $\gamma$  closer to 1.

Further, in Scenario 2 the adjusted  $p$ -values for Endpoints S1 and S2 steadily increase with  $\gamma$  because only one primary test is significant and, as a result, an increasingly smaller fraction of  $\alpha$  is carried over to the secondary analyses. In this case, it will be desirable to choose a smaller value of the truncation parameter to improve the overall probability of success for the primary and secondary endpoints.

To summarize, the truncation parameter serves as a leverage that balances the power functions of the primary and secondary analyses. If the effect sizes of the primary endpoints are uniformly large, a truncation parameter near

1 will help improve the overall power. On the other hand, if the effect sizes are expected to vary across the endpoints, the overall power is likely to be maximized when the truncation parameter is small or in the middle of the  $(0, 1)$  interval. In general, an optimal value of  $\gamma$  can be selected via simulations by maximizing an appropriately defined power function, e.g., the probability of rejecting all false hypotheses or at least one false hypothesis, under realistic assumptions about the effect sizes of individual endpoints.

## 5.5 Tree gatekeeping procedures

The tree gatekeeping methods serve as a unified framework that includes serial and parallel methods as well as a combination of serial and parallel methods with logical restrictions. This framework is quite general and can be used to address multiplicity issues in a wide variety of clinical trial applications.

### 5.5.1 General tree gatekeeping framework

Within the tree gatekeeping framework, gatekeepers are defined at the hypothesis rather than family level, i.e., a hypothesis in a certain family may be *testable* whereas another hypothesis in the same family may not. To give a formal definition, consider a hypothesis in  $F_i$ , say,  $H_{ij}$ , and define two sets of hypotheses associated with  $H_{ij}$  ( $i = 2, \dots, m, j = 1, \dots, n_i$ ). The sets are denoted by  $R_{ij}^S$  (serial rejection set) and  $R_{ij}^P$  (parallel rejection set). These sets include hypotheses from  $F_1, \dots, F_{i-1}$ , at least one of them is non-empty and, without loss of generality,  $R_{ij}^S$  and  $R_{ij}^P$  do not overlap. The hypothesis  $H_{ij}$  is testable if all hypotheses are rejected in  $R_{ij}^S$  and at least one hypothesis is rejected in  $R_{ij}^P$ , i.e., if the following two conditions hold,

$$\max_{k,l \in R_{ij}^S} \tilde{p}_{kl} \leq \alpha \text{ and } \min_{k,l \in R_{ij}^P} \tilde{p}_{kl} \leq \alpha.$$

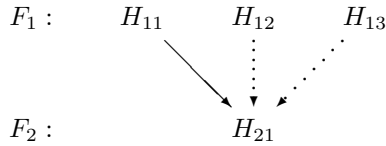
As an example, consider the two-family problem depicted in Figure 5.9. The first family,  $F_1$ , includes three hypotheses ( $H_{11}, H_{12}, H_{13}$ ) and the second one,  $F_2$ , contains a single hypothesis ( $H_{21}$ ). The serial and parallel rejection sets for  $H_{21}$  are defined as follows:

$$R_{21}^S = \{H_{11}\} \text{ and } R_{21}^P = \{H_{12}, H_{13}\}.$$

The hypothesis  $H_{21}$  can be tested only if there is a significant result in  $R_{21}^S$  and at least one significant result in  $R_{21}^P$ .

As was mentioned above, the tree gatekeeping framework includes the serial and parallel gatekeeping frameworks as special cases. Tree gatekeeping procedures simplifies to serial gatekeeping procedures if  $R_{ij}^S = F_{i-1}$  and  $R_{ij}^P$





**FIGURE 5.9:** Tree gatekeeping procedure in a two-family problem. A solid line is used to define a “serial” connection and dotted lines are used for “parallel” connections.

is empty for all hypotheses  $H_{ij}$ ,  $i = 2, \dots, m$ , and to parallel gatekeeping procedures if  $R_{ij}^S$  is empty and  $R_{ij}^P = F_{i-1}$  for all hypotheses  $H_{ij}$ ,  $i = 2, \dots, m$ .

The tree gatekeeping methodology was proposed in Dmitrienko, Wiens, Tamhane and Wang (2007) and was motivated by multiple testing problems that arise in trials when decision trees include multiple branches and/or logical restrictions, e.g.,

- Clinical trials with complex hierarchically ordered hypotheses, e.g., hypotheses associated with multiple endpoints (primary, secondary and tertiary) and multiple test types (noninferiority and superiority), e.g., a hypertension clinical trial with multiple endpoints and noninferiority/superiority tests (Dmitrienko et al., 2007, Section 5).
- Dose-finding studies with multiple endpoints and logical restrictions, e.g., a Type II diabetes clinical trial with a primary and two secondary endpoints (Dmitrienko et al., 2006a) and the metformin-rosiglitazone combination therapy trial (Fonseca et al., 2002) that included a comparison of two metformin-rosiglitazone regimens to metformin on several endpoints.

Multiple testing problems of this kind are quite complex and cannot be handled within the more basic serial or parallel gatekeeping frameworks.

## Closure-based tree gatekeeping procedures

Dmitrienko, Wiens, Tamhane and Wang (2007) developed a framework for constructing tree gatekeeping procedures based on the Bonferroni test. Unlike parallel gatekeeping procedures introduced in Section 5.4.2, Bonferroni tree gatekeeping procedures do not, in general, have a straightforward stepwise form. To define a tree gatekeeping procedure, one needs to utilize the closure principle and use a weighted Bonferroni test for each intersection hypothesis in the closed family associated with the  $m$  families of interest. Dmitrienko, Wiens, Tamhane and Wang gave a general algorithm for assigning weights to individual hypotheses that takes into account logical relationships among multiple analyses in a clinical trial. Dmitrienko, Tamhane, Liu and Wiens (2008)

noted that Bonferroni tree gatekeeping procedures based on this algorithm may violate the tree gatekeeping property defined above, e.g., a hypothesis in  $F_i$ ,  $i = 2, \dots, m$ , may be rejected even though some hypotheses are retained in  $R_{ij}^S$  or all hypotheses are retained in  $R_{ij}^P$ . To address this problem, Dmitrienko, Tamhane, Liu and Wiens formulated a monotonicity condition which is sufficient to guarantee the tree gatekeeping property. Dmitrienko, Tamhane and Liu (2008) and Kordzakhia et al. (2008) derived a weight assignment algorithm that satisfies the monotonicity condition. This algorithm is given in the Appendix.

Dmitrienko, Tamhane and Liu (2008) defined a general approach to defining a broad family of tree gatekeeping procedures that includes Bonferroni tree gatekeeping procedures as a special case. This approach is based on combining multiple tests across families of hypotheses and enables clinical trial sponsors to set up powerful procedures that take into account complex logical restrictions. Examples include tree gatekeeping procedures based on the Hochberg or Dunnett tests.

### 5.5.2 Combination-therapy clinical trial example

To illustrate Bonferroni tree gatekeeping procedures, we will return to the clinical trial example described in Section 5.2.3. This example involves six hierarchically ordered null hypotheses grouped into four families.

To be consistent with the notation introduced earlier in this section, the hypotheses and families will be defined as follows:

- Family  $F_1$  includes  $H_{11}$  (noninferiority hypothesis for A versus B).
- Family  $F_2$  includes  $H_{21}$  (superiority hypothesis for A versus B) and  $H_{22}$  (noninferiority hypothesis for A+B versus B).
- Family  $F_3$  includes  $H_{31}$  (superiority hypothesis for A+B versus B) and  $H_{32}$  (noninferiority hypothesis for A+B versus A).
- Family  $F_4$  includes  $H_{41}$  (superiority hypothesis for A+B versus A).

Now, to account for the logical restrictions among the six hypotheses (the restrictions are displayed in [Figure 5.3](#)), the serial rejection sets are given by

$$\begin{aligned} R_{21}^S &= R_{22}^S = \{H_{11}\}, \\ R_{31}^S &= R_{32}^S = \{H_{22}\}, \\ R_{41}^S &= \{H_{32}\}. \end{aligned}$$

and the parallel rejection sets are empty.

A Bonferroni tree gatekeeping procedure based on the algorithm defined in the Appendix will be used to control the FWER at the two-sided 0.05 level. The adjusted  $p$ -values produced by this tree gatekeeping procedure are listed in [Table 5.6](#). The table shows that the very first hypothesis,  $H_{11}$ , is rejected at

**TABLE 5.6:** Bonferroni tree gatekeeping procedure in the combination-therapy clinical trial example. The asterisk identifies the adjusted  $p$ -values that are significant at the two-sided 0.05 level.

Family	Hypothesis	Raw $p$ -value	Adjusted $p$ -value
$F_1$	$H_{11}$	0.011	0.011*
$F_2$	$H_{21}$	0.023	0.046*
$F_2$	$H_{22}$	0.006	0.012*
$F_3$	$H_{31}$	0.018	0.046*
$F_3$	$H_{32}$	0.042	0.084
$F_4$	$H_{41}$	0.088	0.088

the two-sided 0.05 level and thus the hypotheses  $H_{21}$  and  $H_{22}$  become testable. Both of them are also rejected and, since  $H_{22}$  is included in the serial rejection sets of the hypotheses in  $F_3$ , the tree gatekeeping procedure tests  $H_{31}$  and  $H_{32}$  at the next step. The adjusted  $p$ -value for  $H_{31}$  is significant but the adjusted  $p$ -value for  $H_{32}$  is not. Since the hypothesis  $H_{41}$  depends on  $H_{32}$ , the former is retained without testing. It can be seen from Table 5.6 that the adjusted  $p$ -value for  $H_{41}$  is greater than 0.05.

It is worth noting that the adjusted  $p$ -values displayed in Table 5.6 are equal to those computed in Dmitrienko and Tamhane (2007, Table IV) even though the latter set of adjusted  $p$ -values was obtained using another method (the method defined in Dmitrienko, Wiens, Tamhane and Wang, 2007). The two methods for implementing Bonferroni tree gatekeeping procedures are based on two different algorithms but they often produce identical sets of adjusted  $p$ -values.

---

## 5.6 Software implementation

This section describes the SAS programs that were used in this chapter to implement serial, parallel and tree gatekeeping procedures. These programs can be downloaded from the book's Web site (<http://www.multxpert.com>).

- Serial gatekeeping procedures. Program 5.1 implements the direct-calculation algorithm defined in Section 5.4.2 to compute adjusted  $p$ -values for the three-branch serial gatekeeping procedure in the Type II diabetes clinical trial example (Section 5.3.3).
- Parallel gatekeeping procedures. Program 5.2 computes adjusted  $p$ -values for the two-stage parallel gatekeeping procedure based on the truncated and regular Holm tests in the cardiovascular clinical trial ex-

ample (Section 5.4.3). This program also utilizes the direct-calculation algorithm.

- Tree gatekeeping procedures. Program 5.3 calculates adjusted  $p$ -values for the Bonferroni tree gatekeeping procedure in the combination-therapy clinical trial example (Section 5.5.2).

## Acknowledgements

Ajit C. Tamhane's research was supported by grants from the National Heart, Lung and Blood Institute.

## Appendix

### Error rate function of the truncated Hochberg test

To compute the error rate function of the truncated Hochberg test for  $|I| > 0$ , note that  $e(I) = 1 - P(a_1, \dots, a_k)$ , where  $k = |I|$ ,

$$\begin{aligned} a_i &= \left( \frac{\gamma}{k-i+1} + \frac{1-\gamma}{n} \right) \alpha, \quad i = 1, \dots, k, \\ P(a_1, \dots, a_k) &= P(U_{(i)} > a_i \text{ for all } i = 1, \dots, k) \end{aligned}$$

and  $U_{(1)} < \dots < U_{(k)}$  are the order statistics of i.i.d. observations from a uniform  $(0, 1)$  distribution. Sen (1999) developed a recursive formula for computing  $P(a_1, \dots, a_k)$ . Using this formula, it can be shown that

$$\begin{aligned} P(a_1) &= 1 - a_1, \\ P(a_1, a_2) &= (1 - a_2)(1 - 2a_1 + a_2) \\ P(a_1, a_2, a_3) &= (1 - a_3)(1 - 3a_1 + a_3 - 3a_2^2 + 6a_1a_2 - 3a_1a_3 + a_3^2). \end{aligned}$$

### Weight assignment algorithm for Bonferroni tree gatekeeping procedures

Assuming the multiple testing problem formulated in Section 5.5.1, consider the closed family associated with the  $n$  null hypotheses in Families  $F_1, \dots, F_m$ . For each intersection hypothesis  $H$ , define the indicator functions  $\delta_{ij}(H)$  and  $\xi_{ij}(H)$  as follows. Let  $\delta_{ij}(H) = 1$  if  $H$  contains  $H_{ij}$  and 0 otherwise,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ . Further, for  $i = 2, \dots, m$  and  $j = 1, \dots, n_i$ , let  $\xi_{ij}(H) = 0$  if  $H$  contains at least one hypothesis from  $R_{ij}^S$  or all hypotheses from  $R_{ij}^P$ . Otherwise, let  $\xi_{ij}(H) = 1$ . A Bonferroni tree gatekeeping procedure is defined by specifying a weighted Bonferroni test for each intersection hypothesis  $H$ . To accomplish this, it is sufficient to set up an  $n$ -dimensional

weight vector for  $H$  denoted by  $v_{ij}(H)$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ . The  $p$ -value for  $H$  is given by

$$p_H = \min_{i,j} \frac{p_{ij}}{v_{ij}(H)},$$

where  $p_{ij}$  is the  $p$ -value for  $H_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n_i$ . Note that  $p_{ij}/v_{ij}(H)$  can be set to 1 if  $v_{ij}(H) = 0$ . Based on the closure principle, the adjusted  $p$ -value for  $H_{ij}$  is found by computing the maximum  $p_H$  over all intersection hypotheses containing  $H_{ij}$ .

The weight vector for  $H$  is constructed sequentially by defining  $m$  subvectors

$$(v_{i1}, \dots, v_{in_i}), \quad i = 1, \dots, m,$$

using the algorithm described below (it is assumed in the algorithm that  $0/0 = 0$ ).

**Family  $F_1$ .** Let

$$v_{1j}(H) = v_1^*(H)w_{1j}\delta_{1j}(H), \quad j = 1, \dots, n_1,$$

where  $v_1^*(H) = 1$ , and let  $v_2^*(H)$  denote the remaining weight, i.e.,

$$v_2^*(H) = v_1^*(H) \left( 1 - \sum_{j=1}^{n_1} w_{1j}\delta_{1j}(H) \right).$$

**Family  $F_k$ ,**  $k = 2, \dots, m-1$ . Let

$$v_{kj}(H) = v_k^*(H)w_{kj}\delta_{kj}(H)\xi_{kj}(H), \quad j = 1, \dots, n_k.$$

The remaining weight is given by

$$v_{k+1}^*(H) = v_k^*(H) \left( 1 - \sum_{j=1}^{n_k} w_{kj}\delta_{kj}(H) \right).$$

**Family  $F_m$ .** Let

$$v_{mj}(H) = v_m^*(H)w_{mj}\delta_{mj}(H)\xi_{mj}(H) / \sum_{l=1}^{n_m} w_{ml}\delta_{ml}(H)\xi_{ml}(H),$$

where  $j = 1, \dots, n_m$ .

# Chapter 6

---

## *Adaptive Designs and Confirmatory Hypothesis Testing*

**Willi Maurer**

*Novartis Pharma AG*

**Michael Branson**

*Novartis Pharma AG*

**Martin Posch**

*Medical University of Vienna*

---

### **6.1 Introduction**

A narrowly focused “one hypothesis, one experiment, one decision” approach is often inadequate for the advancement of medical science and practice. Armitage (1991) wondered how statistics might have developed if R. A. Fisher had been employed in medical or industrial research. He suspected that especially experimental design might have developed differently, for agricultural experiments are essentially non-sequential in contrast to the sequential nature of most medical experiments. This difference is amplified by the very special nature of the “units of observations” in clinical trials, the patients, in comparison to those in other scientific investigations, e.g., with respect to blinding, informed consent and compliance in clinical trials. The nature of biomedical research requires stepwise procedures with respect to the sequence of questions to be addressed by separate trials during the four traditional phases of clinical research.

There are obvious reasons for inspecting accumulating information while a clinical trial is in progress. Ethical considerations in studies with human subjects and economic issues, measured in terms of time, money and the number of patients available for future studies, are the most prominent ones. Many of them lead to the classical question of a sequential design: at what point during the course of a study does sufficient evidence accumulate, in favor of or against the test treatment, for discontinuation to be justified?

Statisticians participating in designing clinical trials often are also con-

fronted with questions from their clinical team members like: Why can't we do an interim analysis and, depending on the results, not only stop a trial for proven efficacy or evident futility, but

- stop or delete one or more of the treatment regimens?
- change treatment regimens?
- change inclusion or exclusion criteria?
- change the primary endpoint of efficacy?
- recalculate the sample size?

Such questions arise naturally and should be carefully considered at the planning stage. In the not-so-distant past the common answer to them was: by such interventions the Type I error rate will be altered, estimates will be biased and no valid statistical methods to deal with them appropriately exist. However, in the last decade much progress has been made to devise statistical methods for adaptive designs that

- guarantee robustness against wrong initial assumptions, and
- reduce the number of patients exposed to inefficient or even hazardous treatments and optimize the information content of the trial.

Statistical inferences based on this novel methodology for adaptive designs allows implementation of design adaptations without inflating the Type I error rate. These adaptations may be based on the unblinded data collected up to an interim analysis as well as external information and the adaptation rules do not need to be specified in advance — an indispensable prerequisite to cope with the unexpected.

In early phases of drug development the designs of clinical trials are typically based on little information regarding the true efficacy of the test treatment, its safety profile, the appropriate regimens (dose, formulation, etc.) and the variability of outcome measures. The learning curve can be expected to be steep in these early phases such that during a single clinical trial essential information on properties of the new treatment may become available even though trials in this phase are typically relatively small and of short duration. Based on this information, changes in the trial design may become desirable or even inevitable. In the later phases of drug development the number of open questions is reduced but trials are usually large and long. In this situation it may be possible to address some of the questions with sufficient reliability at an interim analysis allowing to make design adaptation in order to concentrate on the remaining questions. Whereas in early phases of drug development control of the Type I error rate may not be a high priority, it always helps in the interpretation of data. Control of Type I error rate is, however, of utmost importance if adaptive methods are to be applied, for example,

in confirmatory drug development or when combining Phase II and III trials in a combination Phase II/III trial (also known as adaptive seamless trials or confirmatory stagewise adaptive trials). The reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design (CHMP, 2007) stressed this point and in addition asked for the availability of correct estimates and confidence intervals for the treatment treatment.

In this chapter we will concentrate on adaptive designs that are confirmatory in nature and use the following definitions that, up to some refinements, are those given by Maca et al. (2006):

- An *operationally seamless design* is a clinical trial design that combines objectives which are traditionally addressed in separate trials into a single stagewise trial. While some adaptations of the design of the second stage may be based on results of the first stage, the trial stages should not differ operationally (i.e., essentially the same case report forms, centers, indication, etc., are used) and it should be defined in a single protocol. For confirmatory inference though only data from the second stage is used.
- An *inferentially adaptive seamless design* is a seamless stagewise trial design that allows adaptation of the design based on interim results. For the final analysis and hypothesis testing data from patients enrolled before and after the adaptation is used. Operational aspects are the same or even more restrictive than for an operationally seamless design.

In this chapter we will discuss primarily statistical issues of multiplicity and solutions for inferentially adaptive seamless designs and call them in short *adaptive seamless designs (ASD)*.

When planning a fixed design trial we base it on the best available information at that time. Such information comprises

- the availability of appropriate assessments and measures for efficacy and safety that are clinically meaningful and able to differentiate treatment effects,
- properties of chosen endpoints for efficacy and safety regarding their variability or of placebo and control group response rates,
- clinically relevant and expected treatment effects of test treatments in comparison to control treatment effects.

The more reliable the information, the more efficiently the trial can be planned. Efficiency is meant here in a broad sense as ability to provide the “correct” answers with high probability for a given cost. Wise use of additional information accrued during the trial can only improve the efficiency of the design. In the presence of design adaptations the usual frequentist analyses developed for fixed trial designs need modifications in order to properly ensure control of error rates and bias in estimation.



6.2 Basic principles and methods of error rate control

6.2.1 Introduction

In this section we will first classify the sources of Type I error rate inflation. We will concentrate on those that are occurring in group sequential and adaptive designs in addition to the “classical” ones that may be already present in fixed design trials. General methods and principles for controlling the Type I error rate and confinement of estimation bias due to single and multiple sources of inflation will be given in the sequel.

6.2.2 Causes of multiplicity and bias in adaptive designs

The principal differentiation of adaptive designs compared to traditional fixed designs is the ability to perform interim analyses in order to take decisions affecting the further conduct of the trial. This leads to repeatedly testing of one or multiple hypotheses and the possibility to change design features based on interim data. Since the same interim data is subsequently used for hypothesis testing and estimation such approaches may cause bias in estimation and inflation of the Type I error rate if not adequately controlled. The different sources of bias and basic methods for respective adjustments are listed in Table 6.1. In the sequel we will outline the respective multiplicity issues and basic solutions to tackle them separately as well as in combination.

TABLE 6.1: Sources and control of Type I error rate inflation.

Sources of potential error rate inflation	Techniques for error rate control
Repeated hypothesis testing with early rejection of null hypotheses at an interim analysis.	Classical group sequential designs, e.g., designs based on the $\alpha$ -spending approach.
Adaptation of design and analysis features with combination of information across trial stages, e.g., sample size re-assessment based on the treatment effect, data-driven changes in the timing of the next interim analysis or testing strategy.	Combination of $p$ -values, e.g., the inverse normal method, Fisher’s combination test, conditional rejection probability, adjustment for known adaptation rule.
Multiple hypothesis testing, e.g., adaptive selection among initial hypotheses at an interim analysis.	Classical multiple testing methods, e.g., appropriate tests for intersection hypotheses together with closed procedures.

### 6.2.3 Repeated hypothesis testing at interim analyses in group sequential designs

We review here the first source of Type I error rate inflation listed in Table 6.1 and techniques for error rate control if the other sources listed are absent. Repeated testing of hypotheses occurs in trial designs that foresee interim analyses of accrued data together with formal testing of one or several hypotheses together with the possibility of early rejection or retention of the hypotheses. It is well known that repeated testing of a particular hypothesis (without adjusting the significance level) inflates the Type I error rate. It may also be deflated if interim analyses allow for retention of the null hypothesis and the significance levels of the respective tests are not adjusted. The theory and application of methods to properly plan, conduct and analyze clinical trials allowing for interim testing and stopping is well developed. The design of a trial is called “group sequential” when stopping is only foreseen after having accrued additional data of groups of patients and not just single patients. There exists a vast literature on technical and operational aspects of such trials; see, e.g., the review paper by Emerson (2007) or the books by Jennison and Turnbull (2000), Proschan, Lan and Wittes (2006) and Whitehead (1997). We will introduce here only the basic concepts and notations that are needed in the context of the generalization of group sequential designs to adaptive designs, i.e., designs allowing also interim decisions other than stopping or continuing the trial with an otherwise unaltered design.

In this section we assume that there is only one null hypothesis,  $H$ , to be tested versus an alternative,  $K$ . Emerson (2007) described various possibilities of summary statistics of the accruing data to be compared with thresholds, called stopping boundaries, that determine whether the trial should be stopped (and  $H$  be either rejected or retained) or whether the trial should continue. A pair of hypotheses ( $H$  and  $K$ ) and respective stopping boundaries can always be translated (at least asymptotically) into probabilities of errors “spent” up to a certain interim or the final analysis. We will concentrate here on (asymptotically) standard normally distributed test statistics and the respective “nominal”  $p$ -values, i.e.,  $p$ -values that ignore the opportunities for earlier stopping. Consider a parallel group 2-arm trial with  $k > 0$  planned interim analyses, including the final analysis. The treatment effect in comparison to a control denoted by the single parameter of interest,  $\theta$ , that can take on any real value. The null hypothesis is thus defined by  $H : \theta \leq 0$  and the alternative by  $K : \theta > 0$ . Let  $I_t$  be the amount of statistical information available at interim analysis  $t$ ,  $t = 1, \dots, k$ . In many situations  $I_t$  is proportional to the number of cases (patients) available for the analysis up to interim analysis  $t$  and inversely proportional to the standard deviation of the underlying measure of effect. The respective test statistics, e.g., for comparing a test treatment to a control, taking into account all data up to analysis  $t$  is denoted by  $Z_t$ ,  $t = 1, \dots, k$ . In the most common case, the  $Z_t$ ’s then follow

(asymptotically) a multivariate normal distribution with  $E[Z_t] = \theta\sqrt{I_t}$  and  $\text{Cov}(Z_t, Z_{t'}) = \sqrt{I_t/I_{t'}}$ ,  $1 \leq t \leq t' \leq k$ .

## Stopping boundaries

At the first interim analysis ( $t = 1$ ), the test statistic  $Z_1$  is compared to lower and upper “stopping boundaries”  $l_1$  and  $u_1$ , respectively. If  $Z_1 \leq l_1$ , the trial stops and the null hypothesis  $H$  is retained (or equivalently “futility” is declared). If  $Z_1 \geq u_1$ ,  $H$  is rejected in favor of  $K$  and the trial is also stopped. If  $Z_1 \in (l_1, u_1)$  the trial continues to the next planned interim or final analysis. At the interim analysis  $t$ ,  $t < k$ , this decision process is repeated by comparing the now updated cumulative test statistics  $Z_t$  to the boundaries  $l_t$  and  $u_t$ . If the trial is not stopped at any of the interim analyses, a final test is done with  $Z_k$  being compared to the decision thresholds  $l_k = u_k$ . In this case  $H$  is either retained or rejected. Though stopping for futility is foreseen as an option in almost all group sequential trials, it is not necessarily formally dependent on stopping boundaries, but can be decided upon by an independent data monitoring committee. Such a futility assessment may make use of conditional or predictive probabilities of success and/or emerging trends in data besides of the primary parameters of efficacy, e.g., with regards to safety. In settings where no stopping rule for futility is pre-specified, such stopping cannot inflate the Type I error rate but decreases power. In any case, if the decision rules (boundaries) are pre-specified they need to be defined such that the overall Type I error rate, i.e., the probability to reject  $H$  at any of the interim or at the final analysis, is guaranteed not to exceed a predefined level  $\alpha$ , e.g., one-sided  $\alpha = 0.025$ . If the lower stopping boundaries  $l_t$  (also called “futility boundaries”) are used in simultaneously computing  $(l_t, u_t)$ ,  $t = 1, 2, \dots, k$ , such that the probability to reject the null hypothesis at any of the interim or final analysis is  $\alpha$  (see Equations (6.1) and (6.3)), then the futility boundary is binding.

The probability to reject the null hypothesis at interim analysis  $t$ ,  $t = 1, \dots, k$ , given a true treatment treatment  $\theta$ , is

$$\alpha_t(\theta) = P_\theta \left( \{Z_t \geq u_t\} \cap \bigcap_{s=1}^{t-1} \{l_s < Z_s < u_s\} \right). \quad (6.1)$$

The probability to stop exactly at analysis  $t$  and retain  $H$  can be similarly expressed by

$$\beta_t(\theta) = P_\theta \left( \{Z_t \leq l_t\} \cap \bigcap_{s=1}^{t-1} \{l_s < Z_s < u_s\} \right). \quad (6.2)$$

For  $\theta = 0$ , the probability  $\alpha_t(0)$  is called the  $\alpha$  level spent at interim analysis  $t$ . Under the alternative hypothesis  $K$  with  $\theta = \Delta > 0$ , the probability  $\beta_t(\Delta)$  is called the Type II error rate spent at interim decision  $t$ . Since for

different  $t$ , the events in Equation (6.1) are mutually exclusive the overall Type I error rate  $\alpha(0)$  is

$$\alpha(0) = \sum_{t=1}^k \alpha_t(0), \quad (6.3)$$

and the overall Type II error rate  $\beta_t(\Delta)$  is given by

$$\beta(\Delta) = \sum_{t=1}^k \beta_t(\Delta). \quad (6.4)$$

It should be noted that, if at the final analysis a decision is taken with respect to rejection or retention of  $H$ , i.e., if  $l_k = u_k$ , then  $\alpha(0) = 1 - \beta(0)$ . These spent levels are not to be confused with the “nominal” decision levels. Assuming again that  $Z_t$ ’s follow (asymptotically) a multivariate normal distribution, these nominal levels are  $\alpha_t^* = \Phi(-u_t)$  and  $\beta_t^* = \Phi(-l_t)$  where  $\Phi$  denotes the cumulative distribution function of the standard normal distribution and  $\Phi^{-1}$  its quantile function. These nominal boundaries allow rejection of  $H$  at interim analysis  $t$  and stopping if the observed  $p$ -value,  $p_t = \Phi(-z_t)$ , is smaller than or equal to  $\alpha_t^*$  and suggest retention of  $H$  and stopping if  $p_t$  is larger than  $\beta_t^*$ . Since  $P_0(Z_t \geq u_t) = \alpha_t^*$ , it is easily seen when replacing  $\theta$  by 0 in Equation (6.1) that  $\alpha_t^* \geq \alpha_t(0)$ ,  $t = 1, \dots, k$ . In general, strict inequality holds except for the first interim analysis, where  $\alpha_1^* = \alpha_1(0)$ . We will not go into the details on how boundaries are to be chosen such that they satisfy the Equations (6.1) and (6.3) for a given overall level  $\alpha(0) = \alpha$ . For a given number of interim analyses, there is a large choice of “standard” types of boundaries, ranging from those that make an early rejection relatively difficult (O’Brien-Fleming-type boundaries) to those with equal rejection levels at equally spaced interim analyses (Pocock-type boundaries).

Somewhat more flexibility can be achieved by choosing “spending functions” upfront, first introduced by Lan and DeMets (1983). A spending function is a non-decreasing function  $a(y)$  for  $y \in [0, 1]$  with  $a(0) = 0$  and  $a(1) = \alpha$ , where  $y$  denotes the information time or information fraction, i.e., the cumulative proportion of the (maximum) data available for an interim look. For an analysis done at information time  $y$ , the decision boundaries need to be chosen such that the cumulative  $\alpha$  spent (over the current look and all previous looks) is  $a(y)$ . At all decision points, the boundaries (or equivalently the nominal  $\alpha$ -levels) depend only on the number and timing of looks that have come before, and allow some flexibility in the number and in the timing of the looks that will come after. Due to this additional flexibility they cannot just be looked up in an appropriate table computed by iterative numerical methods, but must be re-computed before each interim analysis by using appropriate software. Such software packages were described and discussed by Wassmer and Vandemeulebroecke (2006). It is important to note, however, that both the fixed information time and the  $\alpha$ -spending approach are invalid if the timing of later looks is based on the observed effect at an interim analysis. For example the timing of the next analysis should not be chosen, using

the observed effect, to optimize the probability of success because a procedure would usually inflate the Type I error rate. If this and other additional flexibility is needed then appropriate methods that allow adaptation of the design and decision rules based on observed effects must be used. These will be introduced in Sections 6.2.4 and 6.2.5.

### 6.2.4 Adaptation of design features: Sample size

At the planning stage of a clinical trial, one of the basic design elements is the sample size. We are using it to illustrate the effect of changes of design features on the operating characteristics of the trial (see second row of [table 6.1](#)). For any trial, whether with a fixed or group sequential design, the sample size targeted will depend on the variability of a measure for between-group differences and the relevant difference  $\Delta$  “to be detected” with power  $1 - \beta$  while controlling the Type I error rate at level  $\alpha$ . Whereas the latter three values can be chosen arbitrarily, the variability is inherently a property of the data to be collected and can only be “guessed” based on historical information from similar trials. In the case of a binary variable, it depends on the incidence rate in the control group and the actual effect difference, in the case of an approximately normally distributed variable on the within-group variances.

Essentially such interim estimates can be based on data where the information on assignments of patients to treatment groups is available (“unblinded interim analysis”) or where this is not the case (“blinded interim analysis”). The consequences of using such interim assessments of variability for updating the final sample size on the Type I error rate is discussed below. Most of the pertinent aspects are already apparent in the basic situation where we have to compare an active treatment ( $j = 2$ ) with a control treatment ( $j = 1$ ) on one key variable  $X$  of interest that is approximately normally distributed. We also assume equal group sample sizes  $n$ . With  $X_{i,j} \sim N(\mu, \sigma_j^2)$  denoting the observation of subject  $i = 1, \dots, n$ , in arm  $j = 1, 2$ , we also assume equal variances  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ . The within-group variance is estimated by

$$s^2 = \frac{s_1^2 + s_2^2}{2}, \text{ where } s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2, \quad j = 1, 2.$$

Here  $x_{i,j}$  denotes the observed value of  $X_{i,j}$  and  $\bar{x}_j$  the mean of  $x_{i,j}$  in group  $j$ . An interim analysis is taken after the data of  $n_1$  patients per group  $j, j = 1, 2$ , is available.

### Sample size adjustment for nuisance parameters

In the next two subsections we discuss basic approaches and properties of sample size adjustment after blinded and unblinded interim analyses for estimation of a nuisance parameter using the simple two-group setting described above.

## Unblinded sample size adjustment

If at the start of a trial we have no information or only unreliable historical data about the variance of the outcome variable on either treatment, it seems most natural to perform an interim assessment of the within group variance. Such a procedure was first suggested by Wittes and Brittain (1990).

On the basis of historical data, a within group variance  $\sigma_0^2$  is assumed for the outcome variable. The true variance in the trial, however, is  $\sigma^2$ . Given  $\alpha$ ,  $\beta$ , the relevant treatment difference  $\Delta$  and  $\sigma_0$ , the sample size per group  $n$  for a one-sided test at level  $\alpha$  is, given by the usual formula

$$n = \lceil 2\{[\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)]\sigma_0/\Delta\}^2 \rceil, \quad (6.5)$$

where  $\lceil \cdot \rceil$  is the ceiling function. At an interim analysis, after the data on  $n_1 = qn$  patients per group are available (e.g., with  $q = 0.5$ ), the within-group variance  $\sigma^2$  is estimated by  $s_1^2$  using the equation above just on the  $n_1 = qn$  patients available per groups each at the interim analysis. If  $s_1^2 \leq \sigma_0^2$ , the trial continues as planned to the total sample size  $2n$ . If  $s_1^2 > \sigma_0^2$ , the sample size is adjusted using the newly estimated variance  $s^2$  instead of  $\sigma_0^2$  in the sample size formula above. If the resulting sample size is larger than the originally foreseen sample size  $n$  per group, it replaces  $n$ , otherwise the size remains  $n$ . At the end of the trial, the data analysis incorporates all observations for testing the null hypothesis of no treatment effect by means of a two sample  $t$ -test and for estimation of the effect difference.

One might assume that the level of the test to be performed at the end of the trial will not be affected by the sample size adjustment since the estimate of the variance is uncorrelated with the estimated difference of treatment effects at the time of the interim analysis. This, however, is not the case. The following heuristic argument is provided to clarify this. Assume that, in fact,  $\sigma^2 = \sigma_0^2$ . If  $s_i^2 > \sigma^2$  at the interim analysis, the second stage of the trial includes more patients than originally planned in the fixed sample size design. Hence the influence of this second part of the trial is larger for the final estimate  $s^2$  of  $\sigma^2$  to be used in the  $t$ -test for testing the hypothesis of no treatment effect. Note that, under the condition  $s_1^2 > \sigma^2$ , the expected value of  $s_1^2$  in the first part is larger than  $\sigma^2$  whereas the expected value of  $s^2$  estimated from the second part remains  $\sigma^2$ . Thus the expected value of the final estimate of  $\sigma^2$  is larger in the fixed sample size case than in the adaptive case. This leads to a more likely rejection of the null hypothesis in the adaptive case than in the fixed sample size case, and hence to an inflation of the Type I error rate beyond  $\alpha$ . The effect, however, is negligible even for moderate sample sizes. For one-sided  $\alpha = 0.025$ ,  $\beta = 0.1$ ,  $\Delta = 1$ ,  $\sigma_0^2 = 2$  and  $q = 0.5$  leading to  $n = 43$  per group and an interim sample size of 22 per group, the true level of the test does not exceed 0.052 for a wide range of true variances  $\sigma^2$ . For smaller sample sizes the excess of true size of the test compared to the envisaged level  $\alpha$ , however, can be substantial. Friede and Kieser (2006) listed various options on how to correct for this bias.

The advantage of such a two-stage adaptive procedure over the fixed sample size approach is essentially robustness of the procedure's power with respect to underestimating the true variance. A disadvantage, however, of sample size adjustment that necessitates the unblinding of the treatment group allocation are the increased efforts necessary to keep the interim results confidential. Especially in confirmatory clinical trials it is important to maintain trial integrity and credibility; any possibility for introducing an operational bias should be avoided (see, e.g., Gallo (2006)). Since the publication of the seminal paper by Wittes and Brittain much progress has been made in the area of sample size re-assessment based on interim estimates of nuisance parameters. It is comprehensively covered in a survey paper by Friede and Kieser (2006). These authors also showed in an earlier paper (Kieser and Friede, 2003) that in many situations estimates of variability that avoid unblinding (i.e., are just based on the pooled data) have comparable or even better performance than those requiring unblinded information. The main results are summarized in the next subsection.

### Blinded sample size adjustment

For sample size adjustments taking into account only an estimate of the variability of the observed treatment difference the following procedures based on an estimate of the pooled variance are recommended: Estimates for the pooled variance that do not need unblinding for adjusting the sample size by replacing  $\sigma^2$  in the usual sample size formula are

$$s_{tot,1}^2 = \frac{1}{2n_1 - 1} \sum_{j=1}^2 \sum_{i=1}^{n_1} (x_{ij} - \bar{x})^2, \quad (6.6)$$

where  $\bar{x}$  is the mean of  $x_{ij}$  over both groups in the first stage, and

$$s_{adj,1}^2 = s_{tot,1}^2 - \frac{n_1}{2(2n_1 - 1)} \Delta^2, \quad (6.7)$$

where  $\Delta$  is an assumed mean effect difference.

The estimate for the true variance  $\sigma^2$ , given by Equation (6.6), is unbiased under the null hypothesis, i.e. if the true treatment difference is 0, otherwise it overestimates  $\sigma^2$ . The second term in Equation (6.7) corrects for this, i.e.,  $s_{adj,1}^2$  is an unbiased estimate of  $\sigma^2$  if the true effect difference is  $\Delta$ .

There are various options when using these estimates for the sample size adjustment. The final sample size per group is usually bounded by a value  $n_{max}$  and there is also the possibility to allow or not to allow the adjusted sample size to be smaller than the originally intended one.

For the unrestricted sample size adjustment Kieser and Friede (2003) show via simulation that under a wide range of assumptions for  $\Delta/\sigma$ ,  $n$  and  $n_1 = qn$ , as well as for different significance levels, the true size of a final  $t$ -test is very close to the given significance level  $\alpha$  (absolute difference less than 0.0001).

The actual power is also very close to the intended one when using  $s_{tot,1}^2$  in the sample size re-calculation and its performance is better than when using  $s_{adj,1}^2$  instead.

Note that these properties hold when a positive treatment difference has to be established. In case of testing for equivalence or non-inferiority, Friede and Kieser (2003) show that the true size of the  $t$ -test using the unadjusted adaptive sample size adjustment above is slightly larger than the aimed for significance level. However, for  $\alpha = 0.05$  it is still smaller than 0.054 if the total sample size is larger than 40.

### Sample size adjustment based on the observed effect size

Of much larger influence on the Type I error rate are design adaptations that depend on the observed effect size at an interim analysis unless design changes are appropriately accounted for. This can most easily be illustrated with a “naive” sample size re-estimation based on interim results as described by Gugerli et al. (1993).

Assume that  $\sigma = \sigma_0$  is known for the following two-stage procedure for testing the null hypothesis  $H : \delta = 0$  versus the two-sided alternative  $K : \delta = \Delta \neq 0$ . Let  $n_1$  be the sample size per group at the interim analysis and  $n_{\max} \leq n$  the largest acceptable sample size per group. After recruitment of  $n_1$  patients the sample mean  $\hat{\delta}_1$  based on the observed values is calculated and replaces the initially assumed difference  $\delta$  to be detected in the ordinary “fixed” sample size (Equation (6.5)) resulting in the re-calculated sample size  $\tilde{N}$ . Since the final sample size  $N$  is restricted by  $n_1 \leq N \leq n_{\max}$ , we have

$$N = \begin{cases} n_1, & \text{if } \tilde{N} \leq n_1, \\ \tilde{N}, & \text{if } n_1 < \tilde{N} \leq n_{\max}, \\ n_{\max}, & \text{if } \tilde{N} > n_{\max}. \end{cases} \quad (6.8)$$

For the second stage of the trial an additional  $(N - n_1)$  patients per group are recruited and the test statistic

$$T_N = \frac{\hat{\delta}_N}{\sigma_0} \sqrt{\frac{N}{2}},$$

is computed, where  $\hat{\delta}_N$  is the observed mean effect difference with  $N$  patients per group.  $H$  is rejected if

$$|T_N| \geq \Phi^{-1}(1 - \alpha/2).$$

The actual Type I error rate of this procedure depends only on  $\alpha$  and  $n_{\max}/n_1$ . For  $\alpha = 0.05$  it increases from 0.05 (for  $n_{\max}/n_1 = 1$ , fixed sample size case) to 0.07998 (for  $n_{\max}/n_1 = \infty$ ). Dependent on the sample size adaptation rule (combined with early retention and rejection of hypotheses) the inflation can be larger than this value. Proschan and Hunsberger (1995) calculated an upper limit for the Type I error rate and showed that in the worst case the actual level



can be more than twice the targeted level  $\alpha$ . This inflation of the significance level clearly must be corrected by an appropriate method. Possibilities are to derive adjusted retention and rejection boundaries which are inherently dependent on the rule for adaptation (as in the aforementioned case where the sample size re-calculation formula and the particular test statistics are pre-specified and binding), see Gugerli et al. (1993), or to exploit general invariance principles that allow adjustments independent of the decision rules or of the procedure used for the adaptation. A simple procedure would be to apply the Bonferroni inequality to the stagewise  $p$ -values computed from the incremental data from disjoint data sets per stage. There are, however, more powerful approaches which are discussed in Sections 6.3 and 6.4.

Adaptation of sample size based on observed treatment differences, however, does not automatically lead to a better performance. The operating characteristics of a design aimed at reassessing sample size based on interim information, obviously, depends on the timing of the interim analysis and the adaptation rule chosen. One option is to adapt the sample size such that the power of the final test, conditional on the information available at the interim analysis, attains a given value. This would include the case where the interim analysis shows that, even under optimistic assumptions, the chance to achieve a statistically significant result is slim and therefore one can consider stopping the trial. It has been shown by Bauer and König (2006) that various versions of conditional power can be highly variable and hence unreliable depending on the amount of data available. An alternative possibility is to influence the sample size via a group sequential approach, e.g by “overpowering” the trial with respect to the originally targeted treatment difference and choosing a spending function such that there is a good chance to stop the trial early if the targeted difference actually applies. Jennison and Turnbull (2003) compared the flexible “variance spending” approach of Fisher (1999) to group sequential approaches with respect to overall power and average sample size and found that the latter are often superior. Tsiatis and Mehta (2003) have shown that, for a given sample size adaptation rule, there is always a group sequential strategy with possibly more interim analyses than the adaptive design that is more powerful than the adaptive design. Brannath et al. (2006), on the other hand, have demonstrated that application of the adaptive conditional error principle (see [Section 6.3.2](#)) is conditionally optimal given the observed data up to an interim analysis.

### **6.2.5 Multiple hypothesis selection and testing in adaptive designs**

The third cause for inflation of Type I error rate listed in [Table 6.1](#) is the simultaneous testing of multiple hypotheses. Situations where more than one hypothesis is to be tested are manifold. This multiplicity of hypotheses and questions to be answered, as described in Chapter 2, arises due to, e.g.,

- multiple treatment arms leading to multiple comparison problems,
- multiple subgroups defined by predictive factors,
- multiple endpoints.

Methods for dealing with multiplicity issues in a non-sequential setting are described in Chapter 2. In the context of adaptive trials all these situations lend themselves to improvements over fixed designs by offering the possibility to use interim information for decision making regarding the respective hypotheses. Such decisions can be

- early rejection of some of the hypotheses,
- early retention of hypotheses or selection of hypotheses to be considered further,
- adaptation of the weights given to the hypotheses and respective parts of the trial aimed at testing the respective hypotheses. This can be achieved, e.g., by changing the sample size ratios of treatment arms or subgroups or by changing the analysis strategy.

The first two decisions ask for a generalization of repeated hypothesis testing in a group sequential trial to more than one hypothesis. The third one requires an extension of methods to handle design adaptations to the case of multiple hypotheses. There are different ways to combine the error rate control with respect to repeated hypothesis testing and the multiplicity of hypotheses. In general, for the latter the closure principle can be used for controlling the familywise error rate (FWER) defined in Section 2.2. A short review of this principle is provided in the next paragraph.

## Closure principle

If several hypotheses are tested in clinical trials it is usually recommended to apply a multiple testing procedure to control the familywise error rate (FWER), i.e., the probability to erroneously reject one or more true null hypotheses. A general principle that guarantees FWER control is the closure principle proposed by Marcus, Peritz and Gabriel (1976) (see Section 2.3.3).

## Combining repeated and multiple testing

A simple means for combining repeated testing of the same hypothesis with the testing of multiple (different) hypotheses, is to use the Bonferroni test to bound the error rate for one of the two sources of multiplicity and use a “standard” approach for the other. Assume that there are  $h$ ,  $h > 1$ , hypotheses  $H_j$ ,  $j \in J_1 = \{1, \dots, h\}$  to be tested in a group sequential trial with interim and final analyses, respectively, at  $t = 1, \dots, k$ .

In the following we discuss three approaches to adjust for this double multiplicity:

- Approach 1. Apply the Bonferroni test to the multiple hypotheses.
- Approach 2. Apply the Bonferroni test to the repeated testing of the different hypotheses.
- Approach 3. Define rejection boundaries and stopping rules for all intersections of the  $h$  elementary hypotheses.

Approach 1 was described, e.g., by Follmann, Proschan and Geller (1994) for the case of pairwise comparison in multi-armed trials or more generally in the book of Jennison and Turnbull (2000). Assign to each of the hypotheses  $H_j$ ,  $j \in J_1$ , a significance level  $\alpha_j$  such that  $\sum_{j=1}^h \alpha_j = \alpha$ , and define univariate testing strategies with appropriate spending functions  $a_j(y)$ ,  $0 \leq y \leq 1$ , separately for each of the  $\alpha_j$ 's as defined and described in Section 6.2.3. Since the probability of erroneously rejecting hypothesis  $H_j$  at an interim or the final analysis is bounded by  $\alpha_j$ , the probability to reject any one of them under the global null hypothesis

$$H_{J_1} = \bigcap_{j=1}^{|J_1|} H_j$$

is bounded by  $\alpha$ . Clearly, the closure principle as described in the previous paragraph can be applied here as well, resulting in a group sequential Holm procedure. As in the non-sequential case, this approach can be very conservative if the test statistics for the  $h$  hypotheses are highly positively correlated.

In Approach 2 we apply the Bonferroni inequality to the repeated hypothesis testing: a set of nominal rejection boundaries  $\alpha_t^*$ ,  $t = 1, \dots, k$ , is fixed such that  $\sum_{t=1}^k \alpha_t^* = \alpha$ . At each interim analysis  $t$ , a multiple testing procedure of level  $\alpha_t^*$  is used for the  $h$  hypotheses. The resulting procedure protects the overall FWER. Note that in this case, due to the positive correlation of the sequential test statistics, the actual multiple levels spent,  $\alpha_t$ ,  $t = 2, \dots, k$ , are smaller than the chosen nominal levels and hence the true overall level is less than  $\alpha$ . The difference between the actual level and  $\alpha$  as well as respective power loss is, however, minor if the nominal levels at the interim analyses are very small (e.g., only one interim analysis with the O'Brien-Fleming-type spending function). If the interim analysis is late, or for other types of spending functions, this approach should be avoided due to a large loss in power.

Approach 3 is based on the closure principle and avoids the power loss caused by the Bonferroni inequality. It has been proposed by Tang and Geller (1999) and assumes that for each intersection hypothesis  $H_I = \bigcap_{j \in I} H_j$  of a subset of the elementary hypotheses  $H_j$ ,  $j \in I \subseteq J_1$ , a test statistics  $T_I$  is available. Furthermore, it is assumed that nominal stopping boundaries can be computed for interim analyses  $t$ ,  $t = 1, \dots, k$ , for a given significance level  $\alpha$ . One version of subsequent decision making proposed by the authors asks for

stopping the trial, say at time  $t^*$ , if the global null hypothesis can be rejected at time  $t^*$  with respect to the nominal rejection boundary. If this happens, a closed testing procedure is performed to identify the elementary hypotheses that can be rejected. Other rejection/stopping strategies are possible, but all depend on the availability of test statistics for the intersection hypotheses and the availability of rejection boundaries for given spending functions. An example of this is the use of linear contrasts for the hypotheses in a particular intersection hypothesis. Implicitly this approach was also used for the case of pairwise comparisons by Follmann, Proschan and Geller (1994). They computed the respective rejection boundaries via simulations for the case of equal variances, equal allocation across arms and looks at equal increments of information time between interim looks.

Irrespective of the approach adopted to adjust for multiplicity, one needs to carefully pre-plan the linkage between early retention or rejection of one or more of the hypotheses at an interim analysis and the impact of stopping all or part (e.g., treatment arms, subgroups, etc) of the trial. No inflation of the Type I error rate is induced, however, if no use is made of potential futility levels in the computation of the rejection boundaries, or if the trial is not stopped, despite the early rejection of one (but not all) of the hypotheses. In the latter case, it must be decided upfront if the result of the last analysis is used for the rejection decision of a particular hypothesis, possibly overruling an earlier rejection.

If some of the initial hypotheses are retained at an interim analysis and this is not exploited in the computation of the rejection boundaries, then all of the above approaches lead to procedures that protect the Type I error rate at level  $\alpha$ . These approaches have one thing in common: the only adaptation based on interim results possible is the early retention of hypotheses. Consider, for example, the case of multiple comparisons of several treatment arms to a control treatment: no adaptation other than dropping of one or several treatment arms, or equivalently selecting treatment arms to be continued after an interim analysis, is allowed. Adaptations such as reallocating the pre-planned total number of second-stage patients to the continued arm are not allowed. Since the stopping for futility is not used in the computation of the interim or final rejection boundaries, the overall Type I error rate is not inflated, irrespective of the reason that led to the selection.

If we assume now that only one interim analysis is planned and no early rejection of one or several of the hypotheses is envisaged, the only adaptation is the selection of treatments to continue into the second stage of the trial. For those treatments that are discontinued, the null hypothesis is retained. Then Approach 1 leads to the Holm procedure at the final analysis (on the combined data from both stages) where the  $p$ -values for the dropped hypotheses are set to 1. Note that if only one active treatment arm and the control arm are continued to the second stage then—to reject the respective null hypothesis—the  $p$ -value for the univariate test of treatment vs. control has to be smaller than

$\alpha/h$ , where  $h$  is the initial number of active treatment arms and respective null hypotheses.

Under the same restrictive assumptions, Approaches 2 and 3 lead to the same procedure, since from an early rejection point of view we have only one stage. One can therefore use any one-stage multiple testing strategy as, for example, the Dunnett test. That is, irrespective of whether the second stage is conducted with one or more of the initial treatments, the final analysis uses the Dunnett multiplicity adjustment based on all  $h = |J_1|$  treatments. It has uniformly more power than the Bonferroni approach, since it exploits the positive correlation between the test statistics. The Bonferroni- and Dunnett-based approaches are discussed and compared to other analysis methods, that allow for other adaptations than just dropping treatment arms, by Bretz et al. (2006) and König et al. (2007).

These latter “truly” adaptive methods will be introduced and illustrated with examples in the next sections.

## Comparison of sample size needed for adaptive vs. separate trial strategies

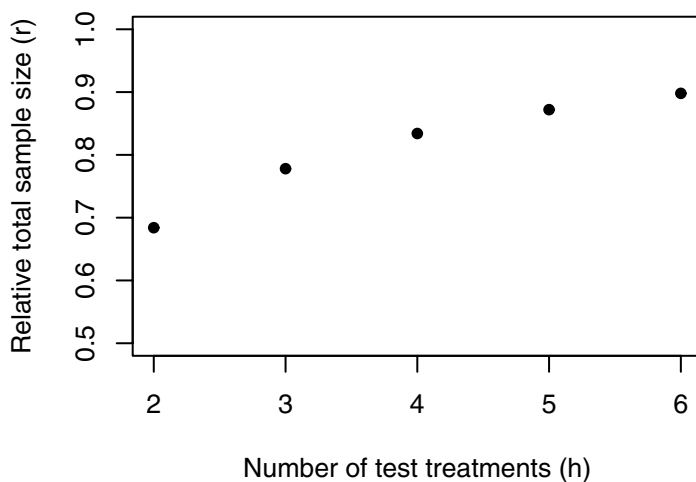
The “one-stage” approach in the case of treatment selection (as defined above) provides a simple means for comparison to a traditional development option with two separate trials. Here the objective of a first trial is the selection of the treatment(s) to be tested in a second, confirmatory trial. The relative sample sizes necessary, given the same power, or the relative power, given the same total sample size, are an important criterion for the choice to be made.

For the adaptive design the “one-stage” Bonferroni procedure is generally more conservative than other adaptive methods which will be described later. But since it is very simple, it lends itself to the derivation of a “rule of thumb” for comparing two-stage adaptive designs with a standard two-trial approach. Such a comparison of the overall sample size given the same power to detect a true treatment effect was described by Bretz et al. (2006). We explain it here in some detail because it helps to understand the reason for gain in power (or equivalently for the reduction in total sample size given the same power).

We consider the case where exactly one out of the initial  $h = |J_1|$  active treatments is selected at an interim analysis. Let  $n_{jt}$  denote the planned sample size (per arm) for treatment  $j = 0, \dots, h$  at stage  $t = 1, 2$ . Let

$$n(\alpha, \beta) = 2\{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\sigma_0/\Delta\}^2$$

denote the sample size that provides a power of  $1 - \beta$  to detect a treatment difference  $\Delta$  with a level  $\alpha$  test in a fixed two-arm comparison. Assume further that in stage 1 the sample sizes per arm  $n_{j1}$  are chosen to be  $n(\alpha, \beta_1)$ . The aim of this stage (or, equivalently, of a separate Phase II study) is primarily to select the right treatment regimens to be investigated further in the second stage (or, equivalently, in an independent confirmatory two-arm trial). The



**FIGURE 6.1:** Relative total sample size  $r$  (Equation (6.9)) of one adaptive versus two separate trials with the same overall power.

sample size is chosen (e.g., via simulations) such that the selected treatment is the “correct” one with a predetermined probability and  $\beta_1$  is then computed post hoc for a given level  $\alpha$ . The latter has no particular meaning in the trial (or stage) and is just chosen to be the same as for the confirmatory test of the selected group in order to simplify the equation. In order to achieve a power  $1 - \beta_2$  for the confirmatory trial in a *classical* design, the necessary sample size is given by  $n(\alpha, \beta_2)$ . The total sample size across both stages and all treatment arms is then

$$n_{class} = (h + 1)n(\alpha, \beta_1) + 2n(\alpha, \beta_2).$$

For the adaptive design with interim treatment selection we assume that the information of both stages is combined using a single one-stage Bonferroni test at the end of the second stage. If we want to achieve the same power  $1 - \beta_2$  for the respective test, the total sample size across both stages for the control group and the continued treatment arm is given by  $n(\alpha/h, \beta_2)$ . The total sample size for the adaptive design is then approximated by  $n_{AD} = (h - 1)n(\alpha, \beta_1) + 2n(\alpha/h, \beta_2)$ . For getting an impression of the values of the ratio of the two sample sizes for different  $h$ , we assume further that the sample sizes per group in the classical design are the same in both trials; i.e., we assume that  $\beta_1 = \beta_2 = \beta$ . Given that the test statistics used for the between

group comparisons are asymptotically normally distributed, we obtain the ratio of sample size of the adaptive vs. that of the classical approach

$$r = \frac{n_{AD}}{n_{class}} = \frac{1}{h+3} \left( h - 1 + 2 \left( \frac{\Phi^{-1}(1 - \alpha/h) + \Phi^{-1}(1 - \beta)}{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)} \right)^2 \right). \quad (6.9)$$

Figure 6.1 shows the relative sample size  $r$  for several values of  $h$  ( $\alpha = 0.025$ ,  $\beta = 0.2$ ), indicating that the classical approach with two independent trials (e.g., a Phase II and a Phase III trial) is less efficient when the number of hypotheses  $h$  is small.

This equation can easily be adapted to the case of the “one-stage” Dunnett procedure by replacing  $\Phi^{-1}(1 - \alpha/h)$  in Equation (6.9) by the respective critical value of the (more powerful) Dunnett test for comparison of  $h$  treatments versus a control with equal sample sizes at level  $\alpha$ . One can see then that the additional reduction in the relative sample size is rather modest (about 1%).

Note that this formula overestimates the true ratio of required sample sizes, since for the adaptive design in reality more powerful adaptive procedures than the “one -stage” procedure may be used. Also the above derivation implicitly assumes that the treatment selection at interim is done randomly and independent of the observed effect sizes at an interim analysis. In practice, however, a treatment with a larger observed effect is more likely to be selected. Conditioned on a selected treatment this increases only the power of the trial with adaptive design but not that of an independent trial. As an example consider the case of 3 treatments vs. a control, assuming that only one of the treatments is effective and that the other two tests have no effect. When selecting the treatment with the best observed effect at interim for continuation to stage two or to a separate new trial, one can find by means of simulation that the actual ratio of sample sizes needed to achieve the same power  $1 - \beta = 0.9$  is  $r^* = 0.695$ , i.e., clearly smaller than the value  $r = 0.762$  given by Equation (6.9). In order to compare the operating characteristics of different design scenarios under various assumptions more reliably, simulations hence are necessary.

---

### 6.3 Principles of adaptive testing procedures

In the preceding section, sources of Type I error rate inflation outlined in Table 6.1 have been reviewed. For the cases of repeated testing of a hypothesis and of multiple hypotheses testing, methods for controlling this error rate separately as well as simultaneously have been reviewed. The effects of other adaptation rules (other than stopping the trial) have been exemplified by sample size reassessment based on observed treatment effect differences. In this section we give an overview on adaptive testing procedures that allow

mid-trial design modifications based on unblinded interim data and possibly external information. Procedures that control the Type I error rate are mentioned in the second row of Table 6.1. One approach is based on combination tests and adheres to a basic invariance principle: separate standardized test statistics are calculated from the samples at different stages and aggregated in a predefined way. Under the null hypothesis the distributions of these separate test statistics are known and essentially invariant under adaptations. Below we introduce adaptive designs based on combination tests (Section 6.3.1) as well as designs based on conditional error rates (Section 6.3.2) which are closely related. These tests control the Type I error rate even if an adaptation rule is not pre-specified *a priori*. We will largely restrict ourselves to the case with just one interim analysis that separates the two stages of the adaptive design.

### 6.3.1 Combination test principle

Assume that a hypothesis  $H$  is tested in a two-stage design. The test decisions are based on  $p$ -values  $p_1$  and  $p_2$  calculated from the separate samples of the two stages. Early decision boundaries are defined for  $p_1$ : If  $p_1 \leq \alpha_1^* < \alpha$ , the trial stops after the interim analysis with an early rejection and, if  $p_1 > \beta_1^* > \alpha$ , it stops with an retention of  $H$  (stopping for futility) <sup>1</sup>. If the trial proceeds to the second stage ( $\alpha_1^* < p_1 \leq \beta_1^*$ ) the decision in the final analysis is based on a combination function  $C(p_1, p_2)$ , which is left continuous and monotonically non-decreasing in both arguments and strictly increasing in at least one: if  $C(p_1, p_2) \leq c$  the null hypothesis is rejected, otherwise retained. Note that for  $\beta_1^* = 1$  no binding stopping for futility is applied. If, in addition,  $\alpha_1^* = 0$  no early rejection is possible. Then, the interim analysis is performed only for adaptation purposes. If, under  $H$ , the  $p$ -values are independently and uniformly distributed on  $[0,1]$  then the level condition for  $\alpha_1^*$  and  $\beta_1^*$  can be written as

$$\alpha_1^* + \int_{\alpha_1^*}^{\beta_1^*} \int_0^1 I\{C(x, y) \leq c\} dy dx = \alpha.$$

Here the indicator function  $I\{\cdot\}$  equals 1 if  $C(p_1, p_2) \leq c$  and 0 otherwise. Actually, it is not necessary to assume that the two  $p$ -values are independent. It suffices to assume that they are  $p$ -clud (Brannath, Posch and Bauer 2002); i.e., the distribution of  $p_1$  and the conditional distribution of  $p_2$  given  $p_1$  are stochastically larger than or equal to the uniform distribution on  $[0,1]$ . If independent sample units are recruited at different stages and tests are applied that control the Type I error probability for any pre-specified significance level  $\alpha$ , the  $p$ -clud condition will apply regardless of the adaptations performed at the interim analyses.

In the sequel we discuss different choices for the combination function

---

<sup>1</sup>To allow a generalization to more than one interim analysis we deviate from the usual terminology for two-stage designs where  $\alpha_1^*$  is denoted by  $\alpha_1$  and  $\beta_1^*$  by  $\alpha_0$ .



$C(p_1, p_2)$ . The first considered was the product of the  $p$ -values

$$C(p_1, p_2) = p_1 p_2,$$

(Bauer, 1989; Bauer and Köhne, 1994). The weighted inverse normal combination function (Lehmacher and Wassmer, 1999; Cui, Hung and Wang, 1999) is defined as

$$C(p_1, p_2) = 1 - \Phi [w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)],$$

with  $0 < w_t < 1$ ,  $t = 1, 2$ ,  $w_1^2 + w_2^2 = 1$ . The weights determine how much importance is given to each stage and may account, e.g., for different pre-planned sample sizes at the two stages. Note, however, that the weights must be pre-planned and may not be adapted during the course of the trial. The inverse normal combination function has a special relation to the group sequential designs discussed in Section 6.2.3. To illustrate this, assume that, under the null hypothesis, the stagewise  $p$ -values  $p_t$ ,  $t = 1, 2$ , are independent and uniformly distributed such that the random variables  $\Phi^{-1}(1 - p_t)$ ,  $t = 1, 2$  are independent and  $N(0, 1)$  distributed. Define  $\tilde{Z}_1 = \Phi^{-1}(1 - p_1)$  and  $\tilde{Z}_2 = w_1 \Phi^{-1}(1 - p_1) + w_2 \Phi^{-1}(1 - p_2)$ . Then, under the null hypothesis, the joint distribution of  $\tilde{Z}_1$  and  $\tilde{Z}_2$  is the same as for the cumulative test statistics  $Z_1$  and  $Z_2$  in a group sequential trial, in which the information times are chosen such that  $\sqrt{I_1/I_2} = w_1$ . Therefore, the critical boundaries,  $\alpha_1^*, \beta_1^*$  and  $c = \alpha_2^*$ , for the inverse normal combination function are identical to the critical boundaries of the group sequential trial and can be computed with standard software for group sequential tests. In the case of just one interim analysis  $\alpha_2^*$  denotes the nominal significance level for the final analysis (see Section 6.2.3).

The analogy between the inverse normal method and group sequential trials extends also to combination tests with more than two stages. Let  $p_t$ ,  $t = 1, \dots, k$ , denote independent and uniformly distributed stagewise  $p$ -values. Define a  $k$ -stage inverse normal combination test by fixing weights  $w_t$ , stopping boundaries  $\alpha_t^*, \beta_t^*$  and combination functions

$$C_t(p_1, p_2, \dots, p_t) = 1 - \Phi \left[ \sum_{s=1}^t \frac{w_s}{\sqrt{\sum_{r=1}^t w_r^2}} \Phi^{-1}(1 - p_s) \right],$$

where  $t = 1, \dots, k$ . The null hypothesis is rejected at stage  $t$  if  $C_t(p_1, \dots, p_t) \leq \alpha_t^*$  and is retained at stage  $t$  if  $C_t(p_1, p_2, \dots, p_t) > \beta_t^*$ . Similar to the two-stage combination test, the critical boundaries  $\alpha_t^*, \beta_t^*$ ,  $t = 1, \dots, k$  can be taken from a group sequential trial with information times  $I_t$  such that  $w_t = \sqrt{(I_t - I_{t-1})/I_k}$ ,  $t = 1, \dots, k$ , setting  $I_0 = 0$ .

### 6.3.2 Conditional error principle

The conditional error function  $A(p_1) = P(\text{reject } H | p_1)$  is defined as the probability to reject the null hypothesis in the final analysis given the first-

stage  $p$ -value  $p_1$ . For a combination test with combination function  $C(p_1, p_2)$ , it is given by

$$A(p_1) = \begin{cases} 1 & \text{if } p_1 \leq \alpha_1^*, \\ \max_{x \in [0,1]} \{x \mid C(p_1, x) \leq c\} & \text{if } \alpha_1^* < p_1 \leq \beta_1^*, \\ 0 & \text{if } \beta_1^* < p_1, \end{cases}$$

(Posch and Bauer, 1999; Wassmer, 1999). Thus, the combination test rejects the null hypothesis in the final analysis if  $p_2 < A(p_1)$ . For example, for the product combination function the conditional error function is given by  $c/p_1$  in case the trial proceeds to the second stage.

An adaptive combination test can also be directly defined via a conditional error function (Proschan and Hunsberger, 1995). Let  $A(p_1)$  denote a monotonic non-decreasing function in  $p_1$  such that

$$\alpha_1^* + \int_{\alpha_1^*}^{\beta_1^*} A(p_1) dp_1 = \alpha.$$

If  $H$  is rejected whenever  $p_1 \leq \alpha_1^*$  or  $\alpha_1^* < p_1 \leq \beta_1^*$  and  $p_2 \leq A(p_1)$ , this procedure controls the Type I error rate.

## Tests based on conditional error rates of standard testing procedures

Müller and Schäfer (2001, 2004) proposed to define the conditional error function based on a pre-planned primary test. Assume that a test with test decision function  $\varphi$  is planned with  $n$  observations to test a null hypothesis  $H$ ; i.e., if  $p$  denotes the  $p$ -value of this hypothesis test, the associated decision function  $\varphi$  of the test performed at level  $\alpha$  is  $\varphi = I\{p \leq \alpha\}$  where  $I$  is the indicator function. Hence  $H$  is rejected (retained) if  $\varphi = 1$  ( $\varphi = 0$ ). Then, after the collection of the first  $n_1$  ( $1 \leq n_1 \leq n$ ) observations, denoted by  $(X_i)_{i=1}^{n_1}$ , one can replace the remainder of the trial by any other test for  $H$  at the (conditional) level of the conditional error function  $E_H(\varphi \mid (X_i)_{i=1}^{n_1})$ . Note that in the general formulation the conditional error function may depend on the first stage data not only through the first-stage  $p$ -value. This approach permits, in principle, the inspection of the data after every observation. If no adaptations are performed then the originally planned test  $\varphi$  is used. If the original test  $\varphi$  is a group sequential design, then early rejection and/or retention of  $H$  can be incorporated also into such trials.

Note that in the presence of nuisance parameters  $E_H(\varphi \mid (X_i)_{i=1}^{n_1})$  is not defined by the null distribution of  $\varphi$  but typically depends on the nuisance parameters and is therefore unknown. For large sample sizes, however, one can work with estimates of the nuisance parameters. For small samples more complex methods have to be used (Posch, Timmesfeld, König and Müller, 2003 and Timmesfeld, Schäfer and Müller, 2007).

## 6.4 Adaptive multiple testing procedures

Assume that a set of  $h$ ,  $h > 1$ , elementary null hypotheses  $H_j$ ,  $j \in J_1$ ,  $J_1 = \{1, \dots, h\}$ , is to be tested familywise error rate at level  $\alpha$ ; i.e., the FWER is controlled at  $\alpha$ , in an adaptive two-stage design. The hypotheses  $H_j$  may correspond to different treatments, subgroups, endpoints or a combination thereof. At an interim analysis a certain subset of these hypotheses is selected, the index set of which is denoted by  $J_2$ ,  $J_2 \subseteq J_1$ . In the second stage, observations are collected for the selected hypotheses only. Note that the selection of hypotheses may be based on unblinded interim data as well as external information. For example, one could select a treatment based on the first-stage efficacy and safety data of the adaptive study as well as from other studies executing in parallel. In the final analysis the elementary hypotheses  $H_j$ ,  $j \in J_1$ , are tested controlling the FWER in the strong sense.

We consider adaptive multiple testing procedures that are based on the application of the closure principle to adaptive tests (Kieser, Bauer and Lehman, 1999; Bauer and Kieser, 1999; Hommel, 2001; Hommel and Kropf, 2001; Posch et al., 2005). For reviews with application to treatment selection, see Bretz et al. (2006, 2009a). In the following we discuss applications of the closure principle to adaptive designs based on combination tests or conditional error rates.

### 6.4.1 Adaptive closed tests

When hypotheses are selected in a two-stage design, no second-stage observations for the dropped elementary hypotheses are available. Let  $J_1$  denote the index set of all initial elementary hypotheses, and  $J_2$  the index set of the elementary hypotheses retained in stage 2. Further, let  $\mathcal{P}(I)$  denote the power set of an index set  $I$ , i.e., the set of all subsets of  $I$ . To define tests for the intersection hypotheses including dropped elementary hypotheses, we have to consider three cases:

- The intersection hypotheses  $H_I$ ,  $I \in \mathcal{J}_A$ , where  $\mathcal{J}_A = \mathcal{P}(J_2)$  where data from both stages is available.
- The intersection hypotheses  $H_I$ ,  $I \in \mathcal{J}_B$ , where  $\mathcal{J}_B = \mathcal{P}(J_1 \setminus J_2)$ , for which only the first-stage data is available.
- The remaining intersection hypotheses  $H_I$ ,  $I \in \mathcal{J}_C$ , where  $\mathcal{J}_C = \mathcal{P}(J_1) \setminus (\mathcal{J}_A \cup \mathcal{J}_B)$ , which are the intersections of a mixture of continued and discontinued elementary hypotheses.

## Closed combination tests

In the planning phase for each hypothesis  $H_I$ ,  $I \in \mathcal{P}(J_1)$ , a combination test  $C_I$  (i.e., a combination function and stopping boundaries) and a first-stage test with  $p$ -value  $p_{I,1}$  are specified. In the interim analysis one can reject all individual hypotheses  $H_j$ ,  $j \in J_1$ , for which all intersection hypotheses  $H_I$ ,  $I \in \mathcal{P}(J_1)$ , with  $j \in I$  can be rejected at the first stage. In contrast, one has to retain all individual hypotheses  $H_j$  for which there exists a hypothesis  $H_I$ ,  $I \in \mathcal{P}(J_1)$  with  $j \in I$  which is retained with its combination test in the interim analysis. After the selection of hypotheses for the second stage, second-stage  $p$ -values for all intersection hypotheses  $H_I$ ,  $I \in \mathcal{P}(J_1)$ , need to be defined:

- For the elementary hypotheses  $H_j$ ,  $j \in J_2$ , second-stage observations are available. Thus, for all intersection hypotheses  $H_I$ ,  $I \in \mathcal{J}_A$ , we can define second-stage  $p$ -values  $p_{I,2}$ .
- For the hypotheses  $H_I$ ,  $I \in \mathcal{J}_B$ , no observations are collected in the second stage and we set  $p_{I,2} = 1$ .
- For all intersection hypotheses  $H_I$ ,  $I \in \mathcal{J}_C$ , we have  $I \cap J_2 \neq \emptyset$  and the second-stage  $p$ -value is defined as

$$p_{I,2} := p_{I \cap J_2,2}, \quad (6.10)$$

where  $p_{I \cap J_2,2}$  is defined in the first bullet point. Note that the set  $I \cap J_2$  depends on the selected treatments and thus on the first-stage data.

In the final analysis the hypothesis  $H_j$ ,  $j \in J_2$ , can be rejected at familywise error rate  $\alpha$  if all intersection hypotheses  $H_I$ ,  $I \in \mathcal{P}(J_1)$ , with  $j \in I$ , are rejected with the respective level  $\alpha$  combination test. According to the closure principle this procedure controls the FWER.

A variety of tests is available for the stagewise tests of intersection hypotheses. For  $I \subseteq J_1$ , denote the number of elementary hypotheses in  $I$  by  $|I|$ . Let  $p_{j,1}$ ,  $j \in J_1$ , and  $p_{j,2}$ ,  $j \in J_2$  denote the unadjusted first- and second-stage  $p$ -values. An adjusted  $p$ -value for  $H_I$  for the Bonferroni test is

$$p_{I,t} = \min[1, |I| \min_{j \in I} p_{j,t}],$$

for the Šidák test is

$$p_{I,t} = 1 - [1 - \min_{j \in I} p_{j,t}]^{|I|},$$

and for the Simes test is

$$p_{I,t} = \min_{j \in I} \frac{|I|}{j} p_{(j),t}, \quad (6.11)$$

where  $I \subseteq J_1$  (for  $t = 1$ ) or  $I \subseteq J_2$  (for  $t = 2$ ) and  $p_{(j),t}$ 's denote the ordered  $p$ -values. For normally distributed test statistics other options to test

the intersection hypotheses are the Dunnett test (Dunnett, 1955) or likelihood ratio tests. In a dose response setting trend tests can be used (Lang, Auterith and Bauer, 2000).

We illustrate the procedure with an adaptive trial incorporating treatment selection. Consider the comparison of three treatments with a control in a parallel group design. Denote the corresponding null hypotheses of no difference in treatment effect between the treatment groups and the control group by  $H_1, H_2, H_3$  such that  $J_1 = \{1, 2, 3\}$ . Assume that a combination function  $C$  has been specified and, for simplicity, no early stopping is planned, i.e.,  $\alpha_1^* = 0, \beta_1^* = 1$ . Let  $p_{1,1}, p_{2,1}$  and  $p_{3,1}$  denote the unadjusted first-stage  $p$ -values for the three treatment-control comparisons and assume that the intersection hypotheses  $H_{123}, H_{12}, H_{13}, H_{23}$  are tested with Šidák tests, such that, e.g.,  $p_{123,1} = 1 - [1 - \min(p_{1,1}, p_{2,1}, p_{3,1})]^3$ . In the interim analysis only treatment 1 is selected, i.e.,  $J_2 = \{1\}$ . The elementary hypotheses  $H_2$  and  $H_3$  have to be retained. Denote the second-stage  $p$ -value for  $H_1$  by  $p_{1,2}$ . Then, according to the above procedure, the second-stage  $p$ -values for the intersection hypotheses  $H_{123}, H_{12}$  and  $H_{13}$  are equal to  $p_{1,2}$ . Consequently,  $H_1$  can be rejected at multiple level  $\alpha$  if  $C(\max(p_{1,1}, p_{12,1}, p_{13,1}, p_{123,1}), p_{1,2}) \leq c$ , where  $c$  is the boundary for the chosen combination function at level  $\alpha_2^* = \alpha$ . For example, for the case of the inverse normal combination function as defined in Section 6.3.1 and no early rejection of a hypothesis, we have  $c = \alpha$ . Note that if  $p_{1,1} \leq \min(p_{2,1}, p_{3,1})$  (i.e., the treatment with the smallest first-stage  $p$ -value is selected) then  $p_{1,1}, p_{12,1}, p_{13,1} \leq p_{123,1}$  such that  $H_1$  can be rejected if  $C(p_{123,1}, p_{1,2}) \leq c$ .

## Closed conditional error tests

In this section we discuss the application of the conditional error rate principle to closed testing. Here, instead of combination tests, for each intersection hypothesis a “classical” test (e.g., the Dunnett test) is used. Then the respective adaptive test based on the conditional error rate of the pre-specified tests is performed.

In the planning phase the total sample size (over both stages) is specified. For all hypotheses  $H_I, I \in \mathcal{P}(J_1)$ , a test  $\varphi_I$  (which is based on the observations from both stages) is defined such that  $E_{H_I}(\varphi_I) = \alpha$ . For the definition of  $\varphi_I$  see Section 6.3.2. Note that the tests  $\varphi_I$  are not tests for the first stage only but for the total pre-planned sample size. One could pre-specify also group sequential tests that allow for early rejection and/or retention of the null hypothesis. For simplicity, we first consider the case where  $\varphi_I$  is a single-stage test. Additionally we assume that no adaptation other than selection of hypotheses is performed. At the end of this section we discuss the extension to group sequential tests and other types of adaptation.

In the interim analysis a set of hypotheses  $H_i, i \in J_2 \subseteq J_1$ , is selected for the second stage. Next we need to define tests for the final analysis for all intersection hypotheses  $H_i, i \in \mathcal{P}(J_1)$ . The (intersection) hypotheses

- $H_I, I \in \mathcal{J}_A$ , are tested with the originally planned tests  $\varphi_I$ .
- $H_I, I \in \mathcal{J}_B$ , are retained.
- $H_I, I \in \mathcal{J}_C$ , are tested with the second-stage tests  $\varphi'_{I \cap J_2}$  at the conditional level (conditional on the interim data) of  $E(\varphi_I | X_1)$ , where  $X_1$  denotes the interim data.

As an example, consider the comparison of two treatments with a control in a parallel group design. We assume a normally distributed endpoint with a known variance and test the (one-sided) hypotheses  $H_1, H_2$  of no difference in means between the treatment and control groups. Let  $n$  denote the pre-planned group sample size across both stages. For the tests of the two elementary hypotheses with decision functions  $\varphi_1$  and  $\varphi_2$ , respectively, we pre-plan  $z$ -tests with  $n$  observations per group. For the intersection hypothesis, we define a Dunnett test. Thus,

$$\begin{aligned}\varphi_1 &= I\{Z_1 > \Phi^{-1}(1 - \alpha)\}, \\ \varphi_2 &= I\{Z_2 > \Phi^{-1}(1 - \alpha)\}, \\ \varphi_{12} &= I\{\max(Z_1, Z_2) > d_\alpha\},\end{aligned}$$

where  $Z_i$  denotes the standardized mean differences between treatment  $i$  and control (based on  $n$  observations per group),  $\Phi^{-1}(1 - \gamma)$  the  $\gamma$ -quantile of the standard normal distribution, and  $d_\alpha$  the critical value of a Dunnett test for two treatments and a control. Assume that after  $n_1$  observations per group an interim analysis is performed. If it is decided to continue with both treatments, the originally planned tests are performed according to the closure principle: The intersection hypothesis is tested with the pre-planned Dunnett test. If this test rejects, the elementary hypotheses are tested with the respective  $z$ -tests. If it is decided to continue only with treatment 1, the second-stage test for the intersection hypothesis  $H_1 \cap H_2$  (taking into account only second-stage data) is the  $z$ -test  $\varphi'_1$  at level  $E(\varphi_{12} | X_{n_1}) = P(\max(Z_1, Z_2) > d_\alpha | X_{n_1})$ . This level can be computed by numerical integration (see König et al., 2008). If the intersection hypothesis can be rejected, the elementary hypothesis  $H_1$  is tested with the pre-planned  $z$ -test (based on  $n$  observations per group).

If the pre-planned tests  $\varphi_I$  are group sequential tests allowing for early stopping, some hypotheses may be rejected or retained at the interim analysis. In this case the conditional error rates of the corresponding hypotheses tests are 1 (in case of early rejection) or 0 (in case of early retention). Note that if a futility bound is crossed for an intersection hypothesis  $H_I$ , all elementary hypotheses  $H_i, i \in I$ , have to be retained.

If other adaptations besides the selection of hypotheses are performed (e.g., sample size adjustment) the intersection hypotheses  $H_I, I \in \mathcal{J}_A$ , cannot be tested with the pre-planned tests anymore. Instead, these hypotheses have to be tested with the second-stage tests  $\varphi'_I$  at conditional level  $E(\varphi_I | X_{n_1})$ .

### 6.4.2 Estimation

Confidence intervals for adaptive designs can be derived based on the duality of hypothesis tests and confidence intervals. In the univariate setting, for each value  $\mu \in \mathbb{R}$ , the null hypothesis  $H(\mu) : \theta = \mu$  is tested against the alternative hypothesis  $K(\mu) : \theta > \mu$  with an adaptive test at level  $\alpha$ , where  $\theta$  is the parameter of interest. Then, a  $100\%(1 - \alpha)$  confidence region is given by all values  $\mu$  that cannot be rejected with the adaptive test (Brannath et al., 2002; Mehta et al., 2007; Brannath, Mehta and Posch, 2007). Similarly, one can construct simultaneous confidence intervals in the multivariate setting (Posch et al., 2005). Assume an adaptive closed combination test of the one-sided hypotheses  $H_j : \theta_j \leq 0$  against  $K_j : \theta_j > 0$ ,  $j = 1, \dots, h$ , has been performed. For simplicity we assume that  $\alpha_1^* = 0$ ,  $\beta_1^* = 1$ ; i.e., no early stopping is performed. To construct confidence intervals for the  $\theta_j$ ,  $j = 1, \dots, h$ , we test for all parameter vectors  $\mu = (\mu_1, \dots, \mu_h)$  the null hypotheses

$$H_j(\mu_j) : \theta_j \leq \mu_j \quad \text{against} \quad K_j(\mu_j) : \theta_j > \mu_j, \quad j = 1, \dots, h, \quad (6.12)$$

and let  $H_{J_1}(\mu)$  denote the corresponding initial global intersection hypotheses. For each  $\mu$  denote the first and second stage  $p$ -value for  $H_{J_1}(\mu)$  by  $p_{J_1,1}(\mu)$  and  $p_{J_2,2}(\mu)$ , respectively, where  $J_2$  denotes again the set of indices of those hypotheses that are continued into stage 2. Now, a  $100\%(1 - \alpha)$  confidence region for  $\theta_j$ ,  $j = 1, \dots, h$  is given by

$$\{\mu \mid C(p_{J_1,1}(\mu), p_{J_2,2}(\mu)) > c\}.$$

Since in general this confidence region is not a cross product of intervals we embed the confidence region into a rectangle. For all  $j \in J_t$  and  $t = 1, 2$ , set

$$p_{j,t}^{\text{adj}}(\mu_j) = \sup_{\xi \in \mathbb{R}^k, \xi_j \leq \mu_j} p_{J_t,t}(\xi). \quad (6.13)$$

For many multiple testing procedures the  $p$ -values  $p_{j,t}^{\text{adj}}$  are of a simple form. Let  $p_{j,t}(\mu_j)$  denote the elementary stage- $t$   $p$ -values for hypotheses (6.12) and assume that  $p_{j,t}(\mu_j)$ ,  $j \in J_1$ , is increasing in  $\mu_j$  and  $\sup_{\mu_j \in \mathbb{R}} p_{j,t}(\mu_j) = 1$ . Then for the Bonferroni test,

$$p_{j,t}^{\text{adj}}(\mu_j) = \min(1, |J_t| p_j(\mu_j))$$

and, for the Šidák test,

$$p_{j,t}^{\text{adj}}(\mu_j) = 1 - [1 - p_j(\mu_j)]^{|J_t|},$$

which are just the classical Bonferroni and Šidák-adjusted  $p$ -values, respectively.

The resulting simultaneous confidence interval for  $\mu_j$ ,  $j \in J_1$ , is given by

$$I_j = \{\mu_j \mid C(p_{j,1}^{\text{adj}}(\mu_j), p_{j,2}^{\text{adj}}(\mu_j)) > c\}. \quad (6.14)$$

If the Bonferroni or Šidák tests are applied to test the intersection hypotheses and only the treatment corresponding to the lowest interim  $p$ -value is selected, the confidence intervals are consistent with the hypothesis tests; i.e., the confidence intervals exclude all parameter values that are rejected by the original multiple testing procedure. However, for other selection rules and multiple testing procedures, the confidence intervals may be inconsistent, that is there are sample points where the confidence interval contains parameter values which are rejected by the multiple test.

To illustrate the procedure, consider the comparison of response rates between three active treatment groups to a control in an adaptive two stage closed test. Define the elementary stagewise  $p$ -values

$$p_{j,t}(\mu) = 1 - \Phi[\sqrt{n_t}(r_{j,t} - r_{0,t} - \mu)/s_{j,t}],$$

where  $s_{j,t} = [r_{j,t}(1 - r_{j,t}) + r_{0,t}(1 - r_{0,t})]^{1/2}$  is the usual estimate of the standard error under the alternative hypothesis and  $r_{j,t}$  denotes the stage  $t = 1, 2$  response rates in the control and treatment groups. Assume that no early stopping boundaries are pre-planned and that for the second stage only the second treatment group (and the control group) is selected. Then, given the intersection hypotheses are tested with Bonferroni tests, the lower confidence bound for the difference in response rates of the second treatment to control is the solution of

$$C\{\min[1, 3p_{2,1}(\mu_2)], p_{2,2}(\mu_2)\} = c \quad (6.15)$$

in  $\mu_2$ , where  $c$  is the critical value at level  $\alpha = 0.025$ . The 50% confidence bound results if  $c$  is replaced by the critical value at level 0.5. See [Section 6.5.1](#) for a numerical example.

Note that for both, the univariate and the multivariate setting, the 50% confidence bounds give conservative point estimates with non-positive median bias. In contrast, besides estimates based only on the second-stage data, no mean unbiased point estimates are available unless a specific adaptation rule is pre-specified. For example, the maximum likelihood estimates computed from the total sample pooled over both stages, are typically mean biased because of data-dependent early stopping, selection and sample size reassessment. However, one can derive upper bounds for this bias (Brannath et al., 2006). If hypotheses are selected in the interim analysis, the bias can be defined in different ways: one can consider the bias of the univariate parameter estimates (univariate bias) for each hypothesis regardless if it was selected or not, or consider only the estimates of the selected hypotheses (selection bias), see Posch et al. (2005). Consider, e.g., the case of treatment selection. Let  $\theta_j$ ,  $j \in J_1$ , denote the mean treatment effect in treatment group  $j$ . Let  $\hat{\theta}_j$  denote an estimate of  $\theta_j$ . Then the univariate mean bias of  $\theta_j$  is given by  $E(\hat{\theta}_j - \theta_j)$  and depends on the selection and sample size reassessment rule, stopping boundaries and parameter vector  $(\theta_j)_{j \in J_1}$ . Note that for the selection rule that selects the treatment with the highest interim effect estimate



(and no other stopping rule or sample size reassessment rule is applied), the univariate bias of the arithmetic mean of the pooled observations over both stages is negative, since sampling is typically stopped for treatments that show low interim effects.

Assuming that only one treatment is selected at the interim analysis, the bias of the estimate of the selected treatment is given by  $E(\hat{\theta}_{j_2} - \theta_{j_2})$ , where  $j_2$  denotes the index of the selected treatment. Note that  $j_2$  is a random variable. For the selection rule that selects the treatment with the highest interim effect estimate the selection bias of the arithmetic mean is typically positive, since only treatments that show a high interim effect are considered. Cohen and Sackrowitz (1989) derived a mean unbiased point estimate for this selection rule for designs with no sample size reassessment and without early stopping for efficacy or futility. However, such bias correction comes at the cost of an increased mean squared error.

---

## 6.5 Case studies

In this section, we illustrate how the mathematical framework discussed in the previous sections of this chapter can be applied to specific adaptive clinical trial designs. For each example, the actual trial design is presented and then critiqued. The critique is intended to discuss how the same goal could have been achieved using alternative statistical techniques, e.g., choice of (multiple) hypothesis testing strategies, highlighting where some of the options exist when defining the statistical framework for an adaptive design. For each of the examples, the series of options discussed are clearly not exhaustive. To fix ideas, we focus on two designs: the first focuses on aspects of *treatment selection* while the second example focuses on *subpopulation selection*. Aspects of these two case studies are also described in Schmidli et al. (2006).

### 6.5.1 Treatment selection design

In clinical drug development selecting the “right dose(s)” is critical. Selecting doses that are too low results in ineffective treatments while choosing doses that are too high could result in unnecessary and undesirable side effects. Once the dose-finding component of any drug development strategy is achieved, the formal evidence of effectiveness is then commonly achieved by having two independent trials, both achieving statistical significance (at the one-sided 2.5% level). Such a development strategy may benefit from leveraging an adaptive seamless design (as one of the two pivotal trials that are usually needed for a new drug application). The formal dose-finding study provides a limited range of doses/regimens for consideration in the confirmatory trial(s); however, uncertainty frequently remains regarding the dose(s) to

select as the clinical doses for registration with the health authorities. In the adaptive seamless design, the final dose-selection is achieved using an interim analysis while the statistical inference framework assures statistical integrity and the strict control of the FWER.

The specific example is based on a planned clinical trial (see Posch et al., 2005), using a two-stage adaptive design. At the first stage one of three active treatments  $J_1 = \{1, 2, 3\}$  is selected and at the second stage, the selected treatment is compared to placebo, combining the data from both stages at the multiple level  $\alpha = 2.5\%$ ; i.e., this is an “inferentially” adaptive seamless design (see Section 6.1). The clinical endpoint of interest is binary and the efficient score test statistic for the log odds ratio (following Cox and Hinkley, 1974) is used to test the hypotheses  $H_j : \theta_j \leq \theta_0$  versus  $K_j : \theta_j > \theta_0$ , with  $\theta_0$  denoting the placebo response rate and  $\theta_j$  denoting the response rate for treatment  $j \in J_1$ .

Denote the observed number of events (successes) on treatment  $j \in J_1$  for stage  $t = 1, 2$  by  $s_{j,t}$  and assume that the planned sample sizes per treatment group are balanced, denoted by  $n_t$ ,  $t = 1, 2$ , with  $n = n_1 + n_2$ . The corresponding success rates are then given by  $r_{j,t} (= x_{j,t}/n_t)$ . For the placebo group the corresponding terms are denoted by  $x_{0,t}$  and  $r_{0,t} (= x_{0,t}/n_t)$ , respectively. The standardized efficient score statistic  $Z_{j,t}$  comparing treatment  $j \in J_1$  to placebo, at stage  $t = 1, 2$ , is asymptotically normally distributed, and is given by

$$Z_{j,t} = \frac{\sqrt{n_t}(r_{j,t} - r_{0,t})}{\sqrt{2\bar{r}_{j,t}(1 - \bar{r}_{j,t})}}, \text{ where } \bar{r}_{j,t} = (r_{j,t} + r_{0,t})/2. \quad (6.16)$$

Note that the estimator of the standard error of the treatment difference here is unbiased under the null hypothesis and hence differs from the one used in Section 6.4.2. The corresponding stagewise  $p$ -values for the elementary null hypotheses are given by  $p_{j,t} = 1 - \Phi(Z_{j,t})$ ,  $t = 1, 2$ ,  $j \in J_1$ . In this specific trial, the information from the two stages is combined using the weighted inverse normal combination function, with weights  $w_1 = \sqrt{n_1/n}$  and  $w_2 = \sqrt{n_2/n}$ . No hypothesis testing is foreseen at the first stage nor binding futility boundaries; following Equations (6.3) and (6.4), we thus have  $\alpha_1(0) = \alpha_1^*(0) = 0$ ,  $\alpha_2(0) = \alpha_2^*(0) = 0.025$ ,  $\beta_1(0) = \beta_1^*(0) = 1$  and  $\beta_2(0) = \beta_2^*(0) = 0.975$ . Hence, the critical value at the second stage is equal to  $\alpha$  (2.5%). Although no *binding* futility boundary is defined at the first stage, a treatment is considered eligible for selection only if the one-sided  $p$ -value  $p_{j,1}$ ,  $j \in J_1$ , is no greater than 0.15. Thereafter, the treatment with the smallest  $p$ -value is intended to be selected provided there is no overriding safety concern. Obviously, this futility assessment with retention of the null hypothesis does not inflate the Type I error rate (see Section 6.2.3).

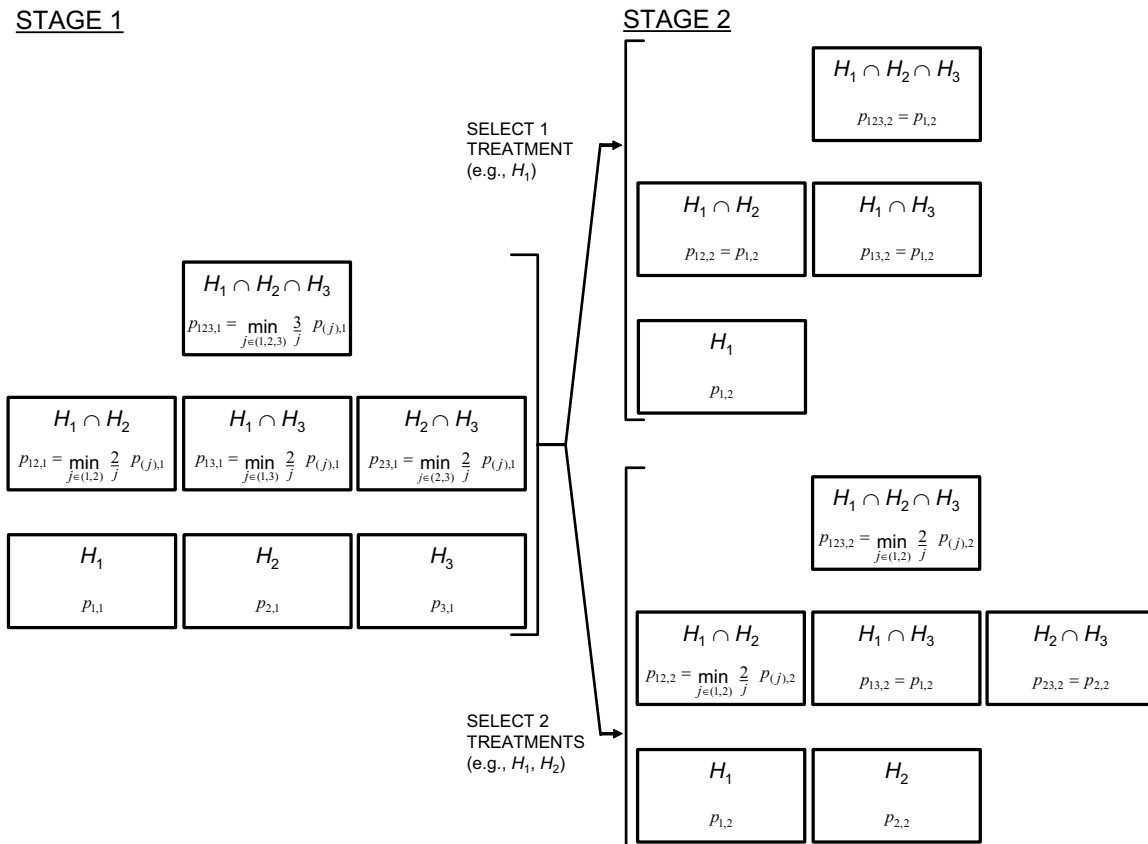
As this design concerns 3 elementary null hypotheses, the closure principle is used to control the FWER. Following Section 6.4.1, the first-stage index set is given by  $J_1$  and the second-stage index set ( $J_2 \subseteq J_1$ ) denotes the subset of treatments selected for continuation to stage 2. For each of the stages, the

intersection hypotheses ( $H_I$ ,  $I \subseteq J_t$ ,  $t = 1, 2$ ) are tested using the Simes test with the corresponding  $p$ -values given by Equation (6.11).

In Figure 6.2, we illustrate how to construct the stagewise  $p$ -values for each intersection hypothesis in the closed test. These  $p$ -values are subsequently plugged into the weighted inverse normal combination function. Although the design is targeted for continuation of only one (the “best”) experimental treatment and placebo at the interim analysis, the case of continuation with two of the three experimental treatments is also provided. Note that when only one experimental treatment is selected at the interim analysis, the second-stage  $p$ -value for each of the tests necessary to complete the closed test is simply the second stage nominal  $p$ -value for the comparison of the selected treatment to placebo. Also, in Figure 6.2 only the definition of the stagewise  $p$ -values is provided for the hypotheses of interest in the second stage. However, more formally, the second-stage  $p$ -values for intersections of null hypotheses that have all been retained are set to 1 (see Section 6.4.1).

In Figure 6.3 we illustrate how this two-stage adaptive design contributes to the clinical development program for this therapy. Specifically, prior to potentially running this study, a dose-finding study is conducted. If only one experimental dose is identified, a classical 2-arm comparative study is conducted. If several ( $\geq 2$ ) candidate doses are identified for further evaluation then the two-stage adaptive seamless design is conducted to refine the treatment selection within the confirmatory trial. In this section we focus on an adaptive design in which three experimental treatment groups have been identified for further investigation in a preceding dose-ranging study. Clearly, if no doses are identified from the dose-finding study, then the follow-up studies are not necessary—this detail path is omitted from Figure 6.3.

As with almost all adaptive designs, and in particular those which additionally use the closed testing procedure, the operating characteristics usually require evaluation using simulation methods. The analysis strategy controls the FWER in the strong sense, by construction. However, there are several aspects and questions for which simulation techniques are (generally) required, e.g., how robust is the design against possible mis-specifications of the assumed treatment differences; how likely is the design to continue under undesirable (e.g., null) treatment response configurations, and so on. In order to evaluate the operating characteristics, several clinical scenarios are evaluated. A subset of the clinical scenarios evaluated for this design are depicted in Table 6.2, expressed as relative differences to placebo (the assumed placebo response rate is 20%). Clearly, many more scenarios are possible and this subset is presented as they represent difficult configurations in terms of power. For this design, we are primarily interested in the probability of selecting treatment 3 and rejecting the corresponding null hypothesis. These clinical scenarios are therefore “difficult” in that the difference between the best treatment (treatment 3) and the next best treatment (treatment 2) ranges from approximately 33% up to approximately 55% of the largest difference to placebo; thus there is a real risk of selecting the wrong treatment group at the end of stage 1.



**FIGURE 6.2:**  $P$ -value definitions of the closed testing procedure for the two-stage adaptive treatment selection design using the Simes test to test intersection hypotheses.

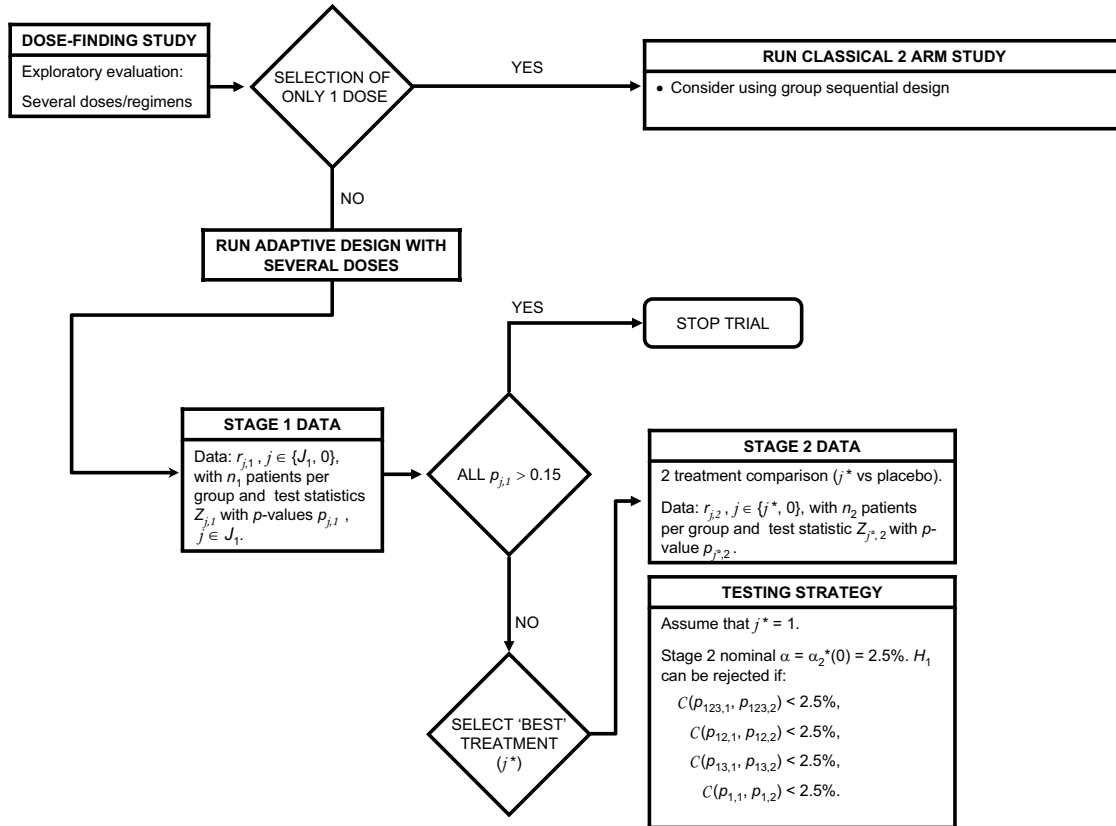


FIGURE 6.3: Decision making and testing strategy for the two-stage adaptive treatment selection design.

**TABLE 6.2:** Response rates relative to placebo in the two-stage adaptive treatment selection design under the null hypothesis (Scenario 0) and three alternative scenarios (Scenarios 1-3).

Treatment	Scenario 0	Scenario 1	Scenario 2	Scenario 3
Placebo	0	0	0	0
Treatment 1	0	0.02	0.02	0.02
Treatment 2	0	0.05	0.05	0.05
Treatment 3	0	0.15	0.12	0.09

**TABLE 6.3:** Design operating characteristics: Treatment selection probability for stage 2, probability to rejecting elementary null hypotheses by the final analysis and expected sample size under the null hypothesis (Scenario 0) and three alternative scenarios (Scenarios 1-3).

Scenario	$P(\text{Selection for Stage 2})$				$P(\text{Rejection of null hypothesis})$				$E(N)$
	Any	Trt 1	Trt 2	Trt 3	Any	Trt 1	Trt 2	Trt 3	
0	0.306	0.107	0.102	0.097	0.021	0.007	0.007	0.007	646
1	0.970	0.006	0.033	0.931	0.920	0.002	0.018	0.901	832
2	0.918	0.019	0.085	0.815	0.765	0.005	0.041	0.720	817
3	0.828	0.044	0.171	0.612	0.520	0.011	0.075	0.434	792

For each clinical scenario, 1,000,000 simulations are used. Posch et al. (2005) assumed that for each simulated trial, the treatment with the highest standardized statistic is selected, provided that the corresponding one-sided  $p$ -value is less than 0.15. If  $p_{j,1} > 0.15$ , for all  $j \in J_1$ , the trial is stopped for futility. The targeted level of power for this design is to achieve 90% probability of selecting treatment 3 and rejecting the corresponding null hypothesis (using the closed testing procedure), under Scenario 1 (see Table 6.2). The search for the stagewise sample size requirements (usually) requires a number of evaluations over a range of viable values (pragmatic, financial, etc). This is also the case for this design. For brevity, the results for 140 patients per group in each stage are given in Table 6.3, however, extensive simulations were run to ensure that this choice of sample sizes per stage generated desirable operating characteristics. In Table 6.3, summary statistics capturing the design characteristics are provided, i.e., the probability of selecting any or a specific treatment group for continuation into stage 2 (column heading ' $P(\text{Selection for Stage 2})$ '), probability of rejecting an elementary null hypothesis at stage 2 having incorporated the futility assessment (column heading ' $P(\text{Rejection of null hypothesis})$ ') and expected sample sizes (column heading ' $E(N)$ ').

Besides comparing different scenarios of adaptive designs and decision rules for subgroup selection, it is always useful to compare the operating characteristics of the chosen adaptive design to those of other development scenarios like a separate Phase II trial for the dose selection and Phase III for confirma-

tion, or just one fixed design Phase III trial with 3 doses. Such comparisons can be found in the papers by Bretz et al. (2005, 2008), Schmidli et al. (2006) or König et al. (2008).

Data analysis: confirmatory hypothesis testing

In this section, following Posch et al. (2005), the hypothesis testing component of the two-stage adaptive treatment selection analysis is illustrated. Aspects of parameter estimation are also demonstrated following Section 6.4.2. Assume observed first-stage response rates of 21% for placebo and 22%, 30% and 36% for treatments 1, 2, and 3, respectively. Furthermore assume that these percentages are based on a total of 140 patients per group. Thus, the first-stage  $p$ -values are  $p_{1,1} = 0.419$ ,  $p_{2,1} = 0.032$  and  $p_{3,1} = 0.0027$ . Given the totality of data up to this interim analysis (both internal to the trial for both efficacy and safety, and up to date external information), assume that treatment 2 is selected for stage 2 based on safety considerations. In the second stage, assume the observed placebo response rate is 19% and for treatment 2, 31% (based on an additional 140 patients per group). The associated  $p$ -value is  $p_{2,2} = 0.0102$ . The definition and how to compute the stagewise  $p$ -values for the respective hypotheses are provided in Figure 6.2. Hence, the stagewise  $p$ -values and corresponding combination (weighted inverse normal) function evaluations are

Stage 1		Stage 2		$C(p_{I,1}, p_{I,2})$
$p_{123,1}$	$= 0.0081 (= 3 \times 0.0027)$	$p_{123,2}$	$= 0.0102$	0.0004
$p_{12,1}$	$= 0.0640 (= 2 \times 0.0320)$	$p_{12,2}$	$= 0.0102$	0.0033
$p_{23,1}$	$= 0.0054 (= 2 \times 0.0027)$	$p_{23,2}$	$= 0.0102$	0.0003
$p_{2,1}$	$= 0.0320$	$p_{2,2}$	$= 0.0102$	0.0016

where  $C(p_{I,1}, p_{I,2}) = 1 - \Phi[(\Phi^{-1}(1 - p_{I,1}) + \Phi^{-1}(1 - p_{I,2}))/\sqrt{2}]$  (i.e., equal weights  $1/\sqrt{2}$  for both stages). Since in our case there are no early stopping boundaries and this combination function is a  $p$ -value itself, the overall  $p$ -value for the combination test is the value of the combination function. Note that all four (intersection) hypotheses can be rejected; i.e.,  $C(p_{I,1}, p_{I,2}) < 0.025$  for all  $I \subseteq J_2$ . It also follows that the multiplicity adjusted  $p$ -value for  $H_2$  is the maximum of the four combination function evaluations (above), i.e., 0.0033. The multiplicity adjusted lower confidence bound for the difference of the response rates of the selected Treatment 2 to placebo—as derived in Section 6.4.2—is 1.6% and is given by the solution of (6.15) with  $C$  denoting the inverse normal combination function. The point estimate with a non-positive median bias is found by setting the critical value  $c$  in (6.15) to 0.5 and is given by 6.1%. For comparison, the unadjusted point estimate is given by 10.5%. If the Šidák test is used instead of the Bonferroni test in the computation of the confidence bound, a less conservative point estimate of 7.8% results.

If rejection boundaries  $\alpha_1(0) \neq 0$  and  $\beta_1(0) \neq 1$  had been defined then, according to Brannath, Posch and Bauer (2002), the overall  $p$ -value for a

normal inverse combination test is given by

$$P(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1(0) \text{ or } p_1 > \beta_1(0) \\ \alpha_1(0) + \int_{\Phi^{-1}(1-\beta_1(0))}^{\Phi^{-1}(1-\alpha_1(0))} \Phi\left(-\frac{\Phi^{-1}[1-C(p_1, p_2)] - w_1 x}{w_2}\right) \phi(x) dx & \text{otherwise,} \end{cases} \quad (6.17)$$

This is based on the ordering of the sample space following Tsiatis et al. (1984).

## Adaptive treatment selection design: a discussion

In this two-stage adaptive treatment selection design, the weighted inverse normal combination function is used, no early stopping is defined (no  $\alpha$ -spending or binding futility boundary) and the multiplicity adjustment (many-to-one comparisons) uses the Simes test within a closed testing approach. We discuss these aspects in reverse order:

## Tests for the intersection hypotheses within the closed test

The Simes test is used to test each of the intersection hypotheses. This approach is valid as there is positive structural correlation due to the many-to-one comparisons (“Dunnett” comparisons). Clearly,

- any other valid level  $\alpha$  test can be used for each of the intersection hypotheses, e.g., (weighted) Bonferroni, Šidák, Dunnett or fixed-sequence tests,
- a mixture of these tests can be used within (and across) stages, e.g., a trend test for the 3-way intersection hypothesis followed by the Dunnett tests for the 2-way intersection hypotheses.

As emphasized in Section 2.6.3, care should be taken when selecting the fixed-sequence approach because, if the first hypothesis cannot be rejected, no further testing is possible. In some settings this may be exactly what is needed; e.g., interest in hypotheses further down the hierarchy is only true if all hypotheses higher in the hierarchy are rejected. A more robust testing strategy against misspecified assumptions is frequently achieved by using tests such as the Bonferroni, Šidák, Dunnett and other tests.

## Decision making at the interim analysis

At the interim analysis, no stopping for rejection (retention) of the null hypotheses is accommodated. Given the prescribed weights  $(w_1, w_2)$  and hence effective information fractions, any valid  $\alpha$ -spending function can be used,



e.g., as available from statistical software packages (see Wassmer and Vandemeulebroecke, 2006). As with classical group sequential designs, it is also possible to explicitly incorporate a futility boundary within the calculations in Equations (6.3) and (6.4). The specific drawback of this approach is that the futility boundary is binding. This is in contrast to the *informal* (non-binding) futility rule implemented in the above two-stage design whereby the FWER is controlled even if the futility rule is violated.

In the planning (and scenario evaluation) phase of this study, it was expected that a total of 140 patients per group/stage would be recruited, with no sample size adaptation. Having specified the weights *a priori*, it was straightforward to determine and adapt the actual number of patients required for the second stage using *unblinded* sample size re-estimation (usually with a practical upper bound), at the end of stage 1. The re-estimated total number of patients were then allocated to those treatment(s) selected for continuation. The adaptive design framework used for this design ensures that such adaptations can be accommodated without inflating the FWER.

Theoretically, the potential scope of adaptations at the end of stage 1 is very broad. For example, the decision at the end of the first stage could be to add an additional interim analysis (Posch and Bauer (2003)) or to add a new treatment arm for continuation into stage 2. Note that, in the context of confirmatory drug development, the addition of a new treatment arm may cause the confirmatory nature of the study to be questioned by a regulatory agency.

Clearly, simulation techniques play an important role in generating the necessary details of the design operating characteristics under various clinical scenarios. In many cases, simulation techniques will be the only way to assess and ensure that the design parameters are tuned (e.g., decision criteria, size of stages, robustness to misspecified treatment effects, etc), delivering an efficient trial design.

## Comparison to the conditional error rate approach

Alternatively, an adaptive treatment selection trial can also be implemented using the conditional error rate approach. Following König et al. (2008), we define for each intersection hypothesis  $H_I$ ,  $I \subseteq J_1$ , a Dunnett test with  $n$  observations per group setting

$$\varphi_I = I\{\max_{j \in I} Z_j \geq d_{|I|, \alpha}\},$$

where  $d_{|I|, \alpha}$  denotes the Dunnett critical values for the comparison of  $|I|$  treatments with a control. Let  $n_1, n_2$  define the pre-planned stagewise sample sizes per group. If in the interim analysis only one treatment is selected (say, treatment 1), in the second stage, a 2-arm comparative study is performed leading to a second-stage  $p$ -value  $p_{1,2}$ . Note that  $p_{1,2}$  is also a  $p$ -value for the intersection hypotheses  $H_{123}, H_{12}, H_{13}$ . Thus, by the closure principle,  $H_1$  can

be rejected if

$$p_{1,2} \leq \min \{E_{H_{123}}(\varphi_{123}|X_{n_1}), E_{H_{12}}(\varphi_{12}|X_{n_1}), E_{H_{13}}(\varphi_{13}|X_{n_1})\}$$

and the pre-planned  $z$ -test  $\varphi_1$  rejects. The conditional error rates can be computed by numerical integration (see König et al., 2008). If the second-stage sample size is chosen adaptively based on the first-stage data, the first-stage test  $\varphi_1$  is replaced by a test that rejects if  $p_{1,2} \leq E_{H_1}(\varphi_1|X_{n_1})$ .

If two treatments are selected at the interim analysis, say, treatments 1 and 2, the second-stage  $p$ -value for  $H_{12}$  (and thus also for  $H_{123}$ ) can be defined as Dunnett adjusted  $p$ -values of a Dunnett test for the comparison of two treatments to a control, which is based on the second-stage observations only.

Simulations suggest that the conditional error approach is in general slightly more powerful than the combination test approach, especially if both treatments are equally effective and more than one treatment is continued to the second stage. However, conditional error rates typically depend on nuisance parameters and only asymptotic expressions are available. Thus, (as in the Dunnett test example assuming the variance is unknown) Type I error rate control is achieved only asymptotically. If, in addition, early stopping boundaries are defined and/or more complex test statistics (e.g., accounting for covariates) are applied, the computation of conditional error rates becomes computationally challenging.

### 6.5.2 Subpopulation selection design

In this section we discuss an adaptive design targeting the selection and confirmation of a treatment benefit within a more targeted subset of patients, i.e., a subpopulation ( $S$ ) among the full population ( $F$ ). Clearly, in this setting, the ability to identify the right patients is pivotal. On the one hand, inappropriately selecting  $S$  when there is benefit in  $F$  results in incorrectly limiting the access to the new therapy. On the other hand, ignoring the (anticipated) existence of  $S$  may culminate in inadvertently missing the treatment benefit as this benefit is too diluted in  $F$  (depending on the prevalence of  $S$ ). The specific example in this section relates to an oncology therapy which was intended as a targeted therapy from its conception.

Historically, developing a more targeted therapy relies upon several serially run clinical trials:

- Step 1. An exploratory hypothesis-generating study using (bio)markers to characterize a subpopulation.
- Step 2. An independent re-affirmation that this subpopulation is indeed sensitive (Phase II study).
- Step 3. A Phase III confirmatory campaign, usually consisting of two separate Phase III studies.

Thus, the traditional approach can be very time consuming and, by construction, is unable to use the cumulative data in support of the final claim of efficacy. In this section we discuss a three-stage adaptive design to make more efficient use of emerging data, combining the objectives of Steps 2 and 3. More specifically, an adaptive interim analysis is used to confirm or disregard  $S$ , with the definition of  $S$  coming from an independent trial (Step 1). This adaptive design has two treatment groups, experimental ( $E$ ) and control ( $C$ ).

The first stage of this adaptive design is the primary decision-making point regarding subpopulation selection. Its objective is to identify if the primary testing strategy should focus on  $S$  alone or both  $F$  and  $S$  (clearly, the latter option is only possible when the decision at the interim analysis is to continue recruiting all patients ( $F$ )) or if the study should be stopped for futility (although no binding futility boundary is defined for this study). The second stage allows for early stopping for superiority with all remaining testing/confirmation undertaken at the end of the third stage, provided that the study has not already been stopped at either the first or second stage. The specific example is based on a planned clinical trial (see Zuber et al., 2006, and Brannath et al., 2007, where the endpoint of interest is a time to event outcome, i.e., progression free survival).

For simplicity, the weighted inverse normal combination approach is used for this design. The treatment effect difference ( $E$  vs  $C$ ) is measured on the (log) hazard ratio scale (denoted by  $\theta$ ) in sub-populations  $S$  and  $S^c$  ( $S^c$  is defined as the complement of  $S$ ). The efficient score test statistics (logrank) are used to test the hypotheses  $H_j : \theta_j \leq 0$  versus  $K_j : \theta_j > 0$ ,  $j \in \mathcal{G}$ , where  $\mathcal{G} = \{F, S\}$ . Care is required when constructing the test statistics for time to event data in the context of adaptive designs to ensure that there is no inflation of the Type I error rate (see Bauer and Posch, 2004). In this study, this issue is avoided by using a stratified logrank statistic based on augmented statistics across  $S$  and  $S^c$  to test  $H_F$ , hence the Type I error rate is not inflated.

For the weighted inverse normal combination function approach, stage  $t$ -specific weights ( $w_t$ ,  $t = 1, 2, 3$ ) need to be specified *a priori* such that  $\sum_t w_t^2 = 1$ , *a priori*. Based on extensive simulations, and specifically for the assumed treatment difference of 23% risk reduction in progression free survival (hazard ratio of 0.77,  $E$  vs  $C$ ), a total of 918 events is required for the final analysis using  $F$  (and 640 events in  $S$ ). Note that there are fewer events in  $S$  due to patients in the first stage recruited in  $F$ . Based on the anticipated follow-up and recruitment patterns, 1,200 patients are expected to be sufficient to complete the study in the targeted timeframe of 3 years. Additionally, the first interim analysis is intended to be based on approximately 170 events (i.e., 18.5% information fraction) and the second interim analysis is expected to take place after approximately 60% information fraction (i.e., 551 events in  $F$  or 384 events in  $S$ ). Therefore,  $w_1 = \sqrt{170/918}$ ,  $w_2 = \sqrt{(551 - 170)/918}$  and  $w_3 = \sqrt{(918 - 551)/918}$ .

As the weighted inverse normal combination method is used, the stopping

boundaries are equal to the classical group sequential boundaries  $(l_t, u_t)$ , with cumulative information fraction  $f_t = \sum_{i \leq t} w_i^2$  at stage  $t$ . See [Section 6.3.1](#) or Lehman and Wassmer (1999) and Wassmer (1999) for more details. For this particular study, the O'Brien-Fleming  $\alpha$ -spending function at a one-sided 0.025 level without a formally defined futility boundary is selected. Therefore, the upper critical values  $(u_1, u_2$  and  $u_3)$  are 4.970, 2.644 and 1.984, respectively, with corresponding nominal  $\alpha$ -levels:  $\alpha_t^*(0) = \Phi^{-1}(-u_t)$ ,  $t = 1, 2, 3$ . It also follows that  $\beta_t^*(0) = 1$  for  $t = 1, 2$  and  $\beta_3^* = 1 - \alpha_3^*(0)$ , following Equations (6.3) and (6.4).

For time to event data, the stagewise  $p$ -values can be based on the increments in the cumulative logrank scores  $(U_{j,t}, t = 1, 2, 3, j \in \mathcal{G})$  at stage  $t$ . Hence

$$p_{j,1} = \Phi(U_{j,1}/\sqrt{V_{j,1}}) \text{ and } p_{j,t} = \Phi\{(U_{j,t} - U_{j,t-1})/\sqrt{V_{j,t} - V_{j,t-1}}\} \text{ for } t = 2, 3,$$

where  $V_{j,t}$  is the observed Fisher's information at stage  $t$  for population  $j \in \mathcal{G}$  (recall that the stratified statistics are utilized for  $j = F$ ).

At the first stage, the intersection hypothesis  $H_{FS}$  is tested using the Simes test, i.e., with multiplicity adjusted  $p$ -value

$$p_{FS,1} = \min\{2 \min(p_{F,1}, p_{S,1}), \max(p_{F,1}, p_{S,1})\}.$$

In terms of the study conduct, it is assumed that at the end of stage 1, early stopping is possible only if all three hypotheses  $(H_{FS}, H_F, H_S)$  can be rejected at the level  $\alpha_1^*(0)$ . This implies a more restrictive stopping criterion for each of the elementary null hypotheses but this rule is primarily motivated by ensuring that adequate data for both  $F$  and  $S$  is available upon study termination. However, as stated earlier, no formal futility boundary is defined for this study. Bayesian posterior probability and predicted power calculations are used to determine which of  $(F, S)$  to select at the end of stage 1. In brief, if the calculated predictive probabilities of a successful (statistically significant) trial in both  $F$  and  $S$  are small, then the trial will be considered for stopping early for futility. If these predictive probabilities are high in both  $F$  and  $S$  then  $S$  will be selected in preference to  $F$  if the corresponding posterior probabilities of treatment benefit in  $S^c$  is low. For further details of these calculations, see Zuber et al. (2006) and Brannath et al. (2009).

If the trial continues with  $F$ , it is assumed that interest resides in both  $H_F$  and  $H_S$ , hence the Simes test is used for the test of the intersection hypothesis, i.e.,

$$p_{FS,t} = \min\{2 \min(p_{F,t}, p_{S,t}), \max(p_{F,t}, p_{S,t})\}, \quad t = 1, 2, 3.$$

Assuming  $S$  is selected at the end of stage 1, the  $p$ -value for all subsequent stagewise intersection hypotheses  $(H_{FS})$  is simply  $p_{S,t}$ ,  $t = 2, 3$ , i.e., the stage-wise  $p$ -value for the elementary null hypothesis  $H_S$ . Furthermore, if at the end of stage 1,  $F$  is selected and only one of the elementary null hypotheses is rejected at the end of stage 2 ( $H_F$  or  $H_S$ ), the study continues and only the

remaining elementary null hypothesis is tested at the end of stage 3. No futility calculations are envisaged at the end of stage 2 analysis. In Figure 6.4 we provide an overview of this three-stage adaptive subpopulation selection design.

The operating characteristics for this design were computed using extensive simulations. As in the case of the adaptive treatment selection design, the analysis strategy controls the FWER, by definition. However, there are many questions needing to be addressed for which simulation methods are necessary, e.g.,

- Question 1. What is the influence of the underlying (unknown) prevalence of  $S$  on the power?
- Question 2. How powerful is the design under different assumed treatment differences in  $S$  and  $S^c$  (and hence  $F$ )?
- Question 3. How does this three-stage design compare with natural competing designs, e.g., classical group sequential approaches?
- Question 4. How robust is the design to selecting the correct (sub)population at the first stage; i.e., how often will the selection algorithm inappropriately select the wrong group for continuation?

It is possible to construct many further questions along these lines; however, for the purpose of illustration, we focus on Questions 1-4 above and present a selected set of simulation results for a range of clinical scenarios from Zuber et al. (2006), where more extensive simulation results can be found. Specifically, for Question 1, prevalence values of 30%, 50% and 80% are assumed in order to evaluate the influence of prevalence on the operating characteristics. Question 2 is assessed by specifying a range of the treatment differences. For this study, the treatment difference in  $(S, S^c)$  is set to 1 or 0.77 (hazard ratio  $E$  vs  $C$ ). For Question 3, the following three competing designs were evaluated:

- Design 1. The three-stage adaptive subpopulation design.
- Design 2. A group sequential design using the same futility criteria as per Design 1 but always recruiting  $F$  and considering both elementary null hypotheses  $(H_F, H_S)$  to be equally important.
- Design 3. A conventional group sequential design that looks only at the full population (i.e.,  $S$  is ignored) but retains the relevant components of the futility criteria of both Designs 1 and 2 that relate only to  $F$ .

Due to the computational intensity of each study design and in comparing competing designs, 1,000 simulations are run, per scenario.

Based on these simulations, if the treatment benefit is assumed to benefit all patients equally and takes the value of the targeted alternative (hazard

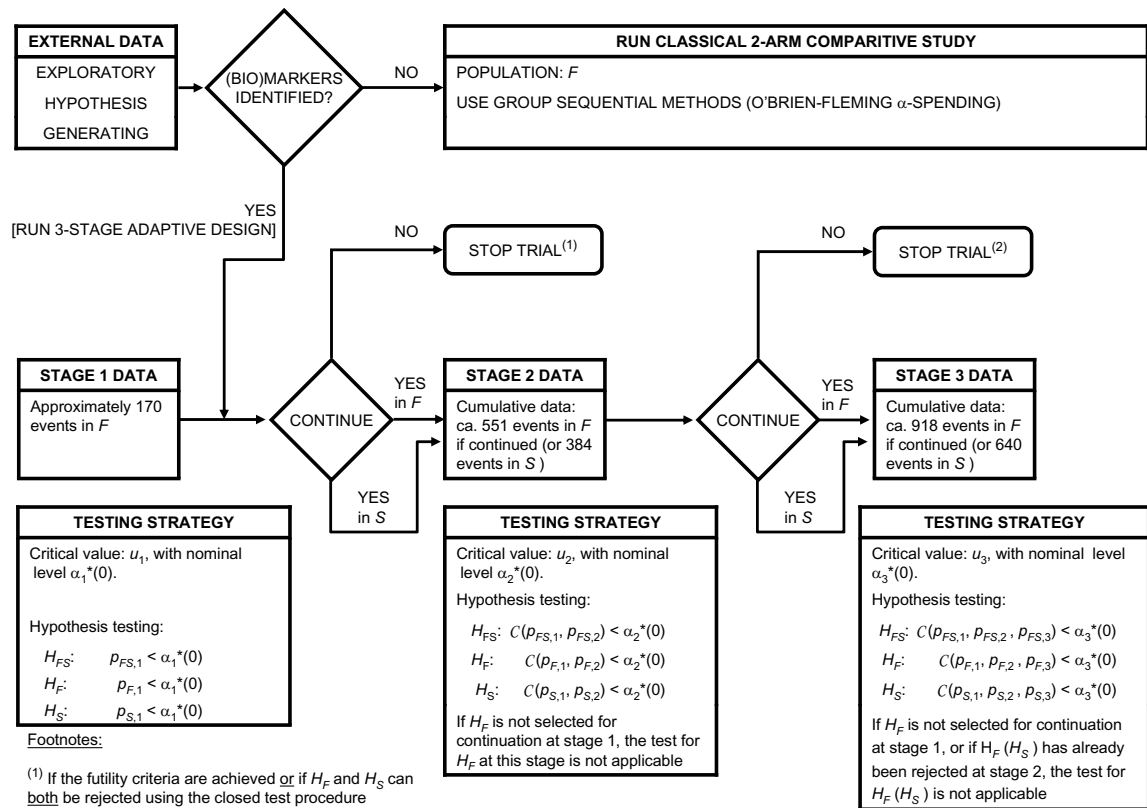


FIGURE 6.4: Decision making and testing strategy for the three-stage adaptive subpopulation selection design.

**TABLE 6.4:** Operating characteristics of the three competing designs in the adaptive subpopulation selection trial.

<i>S</i> prevalence	Effect (hazard ratio <i>E</i> vs <i>C</i> )		Rejection probability		
	<i>S</i>	<i>S<sup>c</sup></i>	Design 1 in <i>S</i> or <i>F</i> (in <i>F</i> )	Design 2 in <i>S</i> or <i>F</i> (in <i>F</i> )	Design 3 in <i>F</i>
30%	0.77	1.00	57% (9%)	39% (14%)	16%
50%	0.77	1.00	71% (24%)	62% (38%)	41%
80%	0.77	1.00	78% (50%)	79% (70%)	73%

ratio = 0.77, *E* vs *C*), Designs 1-3 have approximately equal power in the range of 87-88% (results not shown). Perhaps the more interesting setting is when there is a treatment benefit in *S* alone. In such cases, the adaptive design (Design 1) is as good or better than the other designs (Design 2 and 3). A typical set of simulation results is provided in Table 6.4.

In addressing Question 4, the decisions at the first-stage analysis are (i) stop for futility, (ii) select *F* for continuation or (iii) select *S* for continuation. The simulation results are provided in Table 6.5, corresponding to the futility criteria used to generate the results presented in Table 6.4. More extensive simulation results for a range of futility criteria are presented in Zuber et al. (2006). Clearly, under the global null hypothesis ( $H_{FS}$ ), it is highly desirable to stop at the first interim analysis. Also, if the treatment is benefiting all patients, the most desirable outcome is to continue with the full population (*F*). Least desirable decisions are, e.g., when the treatment is benefiting only *S*; however, there is a high probability to decide to continue with *F*. In Table 6.5, the estimated probabilities for the desirable decisions (given the true state of nature) are boldfaced.

Although the details of the decision criteria are omitted from this chapter for brevity, extensive simulation results similar to those presented in Tables 6.4 and 6.5 were evaluated by Zuber et al. (2006) for a range of decision criteria in order to tune the operating characteristics of the adaptive subpopulation selection design.

**Adaptive subpopulation selection design: a discussion**

Many of the points raised in the discussion of the two-stage adaptive treatment selection design are directly applicable to the three-stage adaptive subpopulation selection design. In particular,

- the choice of test(s) for the intersection hypothesis,
- the ability to incorporate unblinded sample size re-estimation techniques,

**TABLE 6.5:** Probabilities to select different decisions (stop for futility or continue in  $S$  or  $F$ ) at the end of stage 1 in the adaptive subpopulation selection trial.

$S$ prevalence	Effect (hazard ratio $E$ vs $C$ )		Decision probability		
	$S$	$S^c$	Stop for futility and retain $H_{FS}$	Continue in $F$	$S$
30%	1.00	1.00	<b>50%</b>	27%	23%
	0.77	1.00	20%	38%	<b>42%</b>
	0.77	0.77	8%	<b>85%</b>	7%
50%	1.00	1.00	<b>60%</b>	24%	16%
	0.77	1.00	15%	45%	<b>40%</b>
	0.77	0.77	8%	<b>84%</b>	8%
80%	1.00	1.00	<b>65%</b>	26%	9%
	0.77	1.00	15%	59%	<b>26%</b>
	0.77	0.77	10%	<b>81%</b>	9%

- the choice of combination function (and the specific weights for the weight inverse normal method) or conditional error rate approach,
- the timing of the interim analyses,
- the need for extensive simulation techniques to evaluate (competing) design operating characteristics under different clinical scenarios.

One aspect briefly alluded to within the subpopulation selection design relates to the care needed in constructing appropriate test statistics with time-to-event outcomes to avoid inflation of the Type I error rate. In the above example, this is avoided by stratifying the logrank statistic by subpopulations  $S$  and  $S^c$ . More extensive discussion is provided in Bauer and Posch (2004) and Brannath et al. (2008).

## 6.6 Discussion

In this chapter we focused on the different types of multiplicity arising in adaptive designs and methods to properly address the control of the Type I error rate. A short overview on bias in estimation has been given as well. Whereas conservative estimates for the median of variables of interest are easily constructed, this is not the case for the mean. Though proposals for the computation of mean unbiased estimates for certain selection rules exist, very often the bias of, e.g., likelihood estimators is small. It is recommended to investigate size and mean squared error via simulation.

The proposed methods are within the frequentist framework for adaptive



design controlling the FWER by construction. The role of Bayesian approaches for adaptive designs is also increasing with techniques such as those implemented in the ASTIN trial (Berry et al., 2001; Krams et al., 2003) providing an alternative approach to adaptive designs. Bayesian adaptive designs generally require more extensive simulations to ensure that the algorithms and the decision-making criteria adequately control the Type I error rate.

Aside from these statistical challenges the operational and regulatory ones must be also addressed when considering the implementation of an adaptive design. In particular, clinical drug supply and drug packaging will be more demanding, decision makers and the interim/final decision-making process must be determined and the respective trial information must be processed swiftly and reliably for interim reviews and analyses. In order to mitigate the possibility that operational bias might affect the trial results, e.g., by intentional or unintentional dissemination of interim results, maintenance of study integrity is of great importance to ensure the credibility of the results. For adaptive designs to be accepted as confirmatory, an independent Data Monitoring Committee and an analysis team separated from the trial team need to be installed. Professionals directly involved in the conduct of the trial must be blinded with regard to analysis results and adaptation decisions as far as this is possible. These issues and respective solutions were discussed extensively by Quinlan and Krams (2006) and Gallo (2006).

Of additional regulatory concern is the potential heterogeneity of results in the different stages of an adaptive design and the potentially resulting difficulties in providing overall conclusions. The reflection paper on adaptive designs (CHMP, 2007) asked in particular that a strategy for the assessment of homogeneity needs to be planned and described in advance. Adaptive trial designs present many *advantages* over traditional approaches:

- Selection of optimal treatment regimens with smaller patient numbers or with higher reliability.
- Adjustment of ongoing trials to ensure adequate power
- Better informed decision-making.
- Reduced exposure of patients to non-effective doses.
- Reduction of the time between trials or phases of development, leading to a reduction of overall development time.

There are, however, also additional *hurdles* that have to be considered and overcome. They include

- Appropriate statistical methodology is necessary to control the Type I error rate and adjust for bias in estimation.
- Operational complications.
- Regulatory concerns.

Generally the implementation of these designs is more demanding than that of classical fixed designs and careful consideration must be given to the early planning requirements. Statistical modeling and simulation is usually necessary to select among the increased number of design and analysis options, like sample-size stage and treatment arms, spending functions, number of hypotheses to be addressed and respective testing strategies. Operating characteristics must not only be computed for and compared between basic adaptive options and classical approaches, it is also necessary to assess the impact of any interim decision-making on treatment effect estimates to provide guidance to decision makers at interim analyses. Such scenario evaluations will also incorporate—besides of the maximum and expected patient numbers and time and power considerations—the costs in terms of resources needed for planning and conducting the clinical program. Such considerations are considered a crucial part of what the FDA terms “modern protocol design” and help to increase the efficiency of the development process. Adaptive designs may be one of the options in the toolkit to achieve this.

# Chapter 7

---

## *Design and Analysis of Microarray Experiments for Pharmacogenomics*

**Jason C. Hsu**

*The Ohio State University*

**Youlan Rao**

*The Ohio State University*

**Yoonkyung Lee**

*The Ohio State University*

**Jane Chang**

*Bowling Green State University*

**Kristin Bergsteinsdottir**

*Univeristy of Iceland*

**Magnus Karl Magnússon**

*Landspítali-University Hospital*

**Tao Wang**

*Pfizer*

**Eiríkur Steingrímsson**

*Univeristy of Iceland*

---

### **7.1 Potential uses of biomarkers**

A biological marker (a *biomarker* for short) is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention (Biomarkers Definitions Working Group, 2001; FDA 2005b).

Biomarkers can be based on a variety of technologies, and have different uses. Our discussion will be for biomarkers based on microarray technology.

A DNA microarray is a chip with an array of microscopic spots of DNA sequences. They are used to measure relative abundance of nucleic acid sequences in samples. This is done by hybridizing fluorophore-labeled cDNA or cRNA samples to the microarrays, and then measuring the relative intensity of fluorescence emission at each spot.

One use of biomarkers is to determine whether a patient can benefit from a drug. P450s are the major enzymes involved in drug metabolism. AmpliChip CYP450, for example, uses microarray technology to test for variations in the genes that code two of the enzymes (CYP2D6 and CYP2C19) in the P450 family, to see if a patient will have difficulty in metabolizing certain prescription drugs. This is an example of using biomarkers known to measure an important aspect of the biological process, using microarrays as a device.

Another use of biomarker is to predict disease progression. MammaPrint, for example, predicts the likelihood of breast cancer recurrence of a patient using the expression levels of 70 genes. As reported in van't Veer et al. (2002), this 70-gene composite biomarker was developed by measuring the expression levels of approximately 24,000 genes on 78 patients using microarray technology. The 70 genes were selected using multiple testing and machine learning techniques based on error rate control, sensitivity, and specificity considerations. Whether these genes are biologically involved in the disease process, or how they affect disease progression, was not part of the study. MammaPrint itself uses multiplex microarrays that test eight patient samples per array, probing the 70 genes three times per sample. This is an example of developing and using a biomarker based on microarray technology to predict disease progression.

Note that MammaPrint's disease prognosis is given with no reference to any particular treatment. In the next section, we explain how pharmacogenomics can go beyond mere disease prognosis, coupling a drug with a device predicting which patients will respond to the drug.

---

## 7.2 Clinical uses of genetic profiling

Microarrays for clinical use as medical devices are subject to the regulation of the Center for Device and Radiologic Health (CDRH) of the U.S. Food and Drug Administration (FDA). In 2003, the FDA issued its guidance for disease prognostics based on multiplex testing of heritable markers (FDA, 2003). In 2007, Agendia's MammaPrint was approved by CDRH to be marketed as a medical device for breast cancer prognostics. Clinical use of microarray technology is thus a reality. This technology also has potential in drug development, making more efficacious compounds available to patients, with less side effects.

It is well known that most drug development programs fail. However, even if a compound development does not succeed for the entire patient population, there is the possibility that it may still benefit a subgroup of the patient population. In terms of *efficacy*, even when a pharmaceutical trial fails to show sufficient efficacy averaged over the entire patient population, there is the potential that the compound is efficacious for a subgroup of the patient population. In terms of *safety*, even if a compound causes serious adverse events (SAEs) in some patients, there is the possibility that SAEs are confined to a subgroup of the patients. These are the rationales for *pharmacogenomics*.

Analysis of efficacy and safety for certain subgroups based on criteria such as sex and ethnicity is already routinely done. Technology such as microarrays makes it possible to form such subgroups based on the genetic profiles of patients or tissue samples, allowing more refined subgroup analysis in principle.

Pharmacogenomics thus goes beyond mere disease prognosis in that it couples a drug with a device implementing an algorithm predicting which patients will respond to the drug. It is the co-development of a *drug* that targets a subgroup of the patient population, as well as a *device* that can be used to predict whether a patient is in this subgroup of responders to the drug. Since both *drug* and *device* are involved, pharmacogenomics is subject to the joint approval by the Center for Drug Evaluation and Research (CDER) and by the Center for Device and Radiologic Health (CDRH) of the FDA.

In 2005, the FDA issued its Voluntary Genomic Data Submission (VGDS) guidance and drug-diagnostics co-development concept paper (FDA, 2005a, 2005b). With the issuance of these documents, pharmaceutical companies have started banking blood and tissue samples from clinical trials (based on informed consent) for potential pharmacogenomic use.

After highlighting the key statistical issues in these documents in the next section, the rest of the chapter is devoted to a discussion of multiplicity issues in pharmacogenomics.

---

### 7.3 Two stages of pharmacogenomic development

Pharmacogenomic development is a two-stage process. The first stage is to identify a biomarker positive ( $G^+$ ) subgroup of patients for which the compound is extra efficacious, compared to patients in its complement, the biomarker negative ( $G^-$ ) subgroup.

In clinical trials for drug development, efficacy can be defined in terms of higher *average improvement* (over the control group), or in terms of higher *responder rate* (over the control group):

- In Alzheimer's disease trials, efficacy is typically established by comparing mean changes from baseline between treated and control groups.

- In schizophrenia trials, efficacy might be established by comparing mean changes from baseline of the Positive and Negative Syndrome Scale (PANSS) score between treated and control groups. Or it might be established by comparing responder rates between treated and control groups. A responder for a schizophrenia drug might be one who experiences at least a 30% reduction in total PANSS score from baseline.
- For hypertensive drugs, a responder might be one whose systolic blood pressure has been reduced to no more than 120 mm Hg.
- For diabetes drugs, a responder might be one whose HbA1c (glycosylated hemoglobin) is less than 7%.

One might look for a  $G^+$  subgroup based on genotypes already suspected to affect disease outcome. In Alzheimer's disease, for example, one might compare efficacy between carriers and non-carriers of the ApoE "4" allele. Alternatively, one might attempt to discover a  $G^+$  subgroup by comparing the genetic profiles of the responder ( $R^+$ ) patients with the profiles of the non-responder ( $R^-$ ) patients. Using the banked biological samples, measurements on typically a large number of biomarkers are obtained. These marker measurements may be SNP categories obtained from blood samples, or gene expression levels measured from tissue samples, for example. Biomarkers that show substantial differences between the  $R^+$  and  $R^-$  groups are selected, and based on a combination of these selected biomarkers, which might be called a *composite* biomarker or a gene signature, a prognostic classification algorithm is constructed to predict whether a future patient will be a responder or a non-responder to the compound.

At the end of this first stage, the *sensitivity* of the prognostic algorithm, which is the probability that a patient is biomarker positive ( $G^+$ ) given that the patient is a responder, and the *specificity* of the algorithm which is the probability that a patient is biomarker negative ( $G^-$ ) given that the patient is a non-responder, should be estimated. Provided that both the estimated sensitivity and specificity are sufficiently high, pharmacogenomic development proceeds. Otherwise, further pharmacogenomic development is likely to be futile.

In Section 7.4, we discuss multiple testing for differential expressions and for significant composite biomarkers.

If pharmacogenomic development proceeds to the second stage, a new clinical trial is conducted to independently validate the efficacy and safety of the compound for the target subgroup, and to prove that the composite biomarker has sufficient sensitivity and specificity for clinical use.

One issue that is often overlooked is that the process of developing a prognostic device based on gene expressions involves a *change of platform* between the training stage and the validation (and eventual clinical use) stage. As stated in FDA (2005b):

A new test with fewer biomarkers developed for diagnostic purposes (i.e., patient stratification) should be properly validated, ideally in clinical trials that enrolled patients with the intended indication.

Whereas the training study might use microarrays probing many biomarkers, the validation study uses the prognostic chip containing only the genes in the signature that is intended for eventual clinical use. (For example, the validation study of MammaPrint used microarrays probing 70 genes only.)

Another issue that has not been fully addressed is how to design and analyze the training study in order to properly design the validation study. As stated in FDA (2005b), it is important for a pharmacogenomics development plan to be able to compute sample sizes required to meet validation requirements:

When validating a gene or expression pattern, instead of a set of individual biomarkers, a rigorous statistical approach should be used to determine the number of samples, and the methodology used for validation. It is recommended that the validation strategy be discussed in advance with FDA.

In this chapter, we discuss how to statistically design and analyze the training study in order to properly design the validation study. Determination of sample sizes for the validation study are elucidated in Rao, Lee and Hsu (2009).

---

## 7.4 Multiplicity in pharmacogenomics

Two of the sources of multiplicity in pharmacogenomics are multiplicity of individual *biomarkers* and multiplicity of *subgroups*. Individual biomarkers (genes in our discussion) can be selected to form a composite biomarker. Subgroups are defined by the multitude of prediction algorithms that can be formed by the selected individual biomarkers.

To select biomarkers to form a composite biomarker, one can test for the significance of the individual biomarkers controlling a multiple error rate such as the Familywise Error Rate (FWER), generalized Familywise Error Rate (gFWER) and False Discovery Rate (FDR). These error rates are defined in Section 2.2.

Then, after proposing a prognostic algorithm based on the selected biomarkers, one can validate its discriminant power as follows. Each potential prediction algorithm divides the patients into the  $G^+$  or  $G^-$  groups. A patient in the  $G^+$  group is predicted to be a responder, while a patient in the  $G^-$  group is predicted to be a non-responder. Therefore, one can account for the

multiplicity of subgroups by proving that the proposed prognostic algorithm has non-zero discriminant power, even after taking into account the multiplicity of potential prediction algorithms that can be formed by the selected biomarkers.

### 7.4.1 Multiplicity of genes

In the first stage, genes are tested for differential expressions, as having too many genes may hinder construction of an effective classification algorithm and inflate its variability. At this stage, expression levels are measured using microarrays that probe a large number of genes, perhaps the whole genome. For example, the first stage in the training of MammaPrint used microarrays that probe approximately 24,000 genes. With such a large number of genes, if they are tested without adjusting for multiplicity, surely some will be found to be differentially expressed even if genetic makeup has absolutely no bearing on the response. This is the first multiplicity issue in pharmacogenomics. To confidently select genes to train a classification algorithm within the first stage, we control an appropriate multiple testing error rate, in agreement with FDA (2005b):

Statistical considerations in deriving a small number of biomarkers from a large amount of parallel multiplexed data should be properly addressed.

### Two different definitions of Type I error rate

Suppose  $m$  genes are probed in comparing expression profiles between responder and non-responder groups. Let  $\mu_{Ri}$ ,  $\mu_{Ni}$ ,  $i = 1, \dots, m$ , denote the expected (logarithms of) expression level of the  $i$ th gene of a randomly sampled patient from the responder and non-responder group respectively. Let  $\theta_i$  denote the difference of the expected (logarithms of) expression levels of the  $i$ th gene between the two groups,  $\theta_i = \mu_{Ri} - \mu_{Ni}$ .

In the current bioinformatics literature, the (marginal) null hypotheses being tested are

$$H_i : \theta_i = 0, \quad i = 1, \dots, m. \quad (7.1)$$

Multiple testing then generally involves testing  $H_I : \theta_i = 0$  for all  $i \in I$  for  $I \subseteq \{1, \dots, m\}$ .

A Type I error of testing  $H_I$  can be defined two different ways. Let  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$ , and let  $\boldsymbol{\Sigma}$  denote generically all parameters that the observed expression levels depend on (including variances, covariances, skewness, etc.). Let  $\boldsymbol{\theta}^0 = (\theta_1^0, \dots, \theta_m^0)$ , and let  $\boldsymbol{\Sigma}^0$  be the collection of all (unknown) true parameter values.

The traditional definition of the Type I error rate (Casella and Berger, 1990, Definition 8.3.3), is

$$\sup P\{\text{Reject } H_I | \boldsymbol{\theta}, \boldsymbol{\Sigma}\}, \quad (7.2)$$



where the supremum is taken over all possible  $\theta$  and  $\Sigma$  subject to  $\theta_i = 0$ ,  $i \in I$ . In the analysis of gene expression levels, the supremum is taken over all possible joint distributions of expression levels under the null hypothesis (including all possible correlations and expression levels of the other genes with indices not in  $I$ ). The more realistic null hypotheses (which might become more popular as bioinformatics evolves) are defined as

$$H_i : \theta_i \leq \delta \text{ and } H_i : \theta_i \geq -\delta$$

with  $\delta > 0$ . Note that the Type I error rate of a test for  $H_I$  may well depend on the values of  $\theta_j$ ,  $j \notin I$ , in the presence of gene by gene interactions.

A different definition of the Type I error rate for testing  $H_I$ , given in Pollard and van der Laan (2005), for example, is

$$P\{\text{Reject } H_I | \theta^0, \Sigma^0\}. \quad (7.3)$$

Note that  $\theta_i^0 = 0$  for  $i \in I$  under  $H_I$ ,  $\theta_j^0$ ,  $j \notin I$  and  $\Sigma^0$  are unknown. Thus this probability is difficult to compute directly. Pollard and van der Laan's (2005) methods, implemented as the `mtp` function in the `multtest` package in the bioconductor repository of R, estimate  $\theta_j^0$ ,  $j \notin I$  and  $\Sigma^0$  by resampling. Their methods control the Type I error rate *asymptotically* as the number of microarrays goes to infinity.

In practice, statistical decision procedures are applied to *different* studies over time. Therefore, error rate control is more useful if it controls the long run relative frequency of incorrect decisions across different studies. Such a claim is possible provided each test controls the Type I error rate at level  $\alpha$  according to the traditional definition (7.2). See Berger and Wolpert (1988, Pages 71-72) and Berger (1985, Page 23, Page 44). In this sense, perhaps definition (7.2) is more useful than definition (7.3).

## Popular error rates to control in bioinformatics

Let  $V$  denote the number of false rejections and let  $R$  denote the total number of rejections. The two most popular quantities to control in bioinformatics are the *false discovery number*  $V$  and *false discovery proportion*  $V/R$ . The rates of these errors are reported either as an *exceedance probability* (relative frequency of  $V$  or  $V/R$  exceeding a specification across many multiple tests), or an *expectation* (average of  $V$  or  $V/R$  across many multiple tests). Such reporting can either be *unconditional*, averaged over all studies, or *conditional* on the data, restricted to studies with an observed number of total rejections  $R = r$ , for example.

The FWER and gFWER are exceedance probabilities. The FDR is an expectation error rate. The step-up test of Benjamini and Hochberg (1995) is a popular FDR-controlling method.

Another popular method in bioinformatics is the Significance Analysis of Microarrays (SAM) proposed by Tusher, Tibshirani and Chu (2001). Instead

of controlling the FDR, it aims to control  $E(V)/R$ , under the configuration that all the null hypotheses are true.

Reporting of an error rate which is an *expectation* may be inadequate if the quantity being controlled (or a component of it) turns out to be highly variable for a statistical method. Gordon et al. (2007) showed that, in terms of  $V$ , the number of false discoveries, the Benjamini and Hochberg (1995) method is less stable than the Bonferroni method.

If controlling  $V$  guards against incorrect decision-making, and the Type I error rate is controlled according to the traditional definition (7.2), then FWER or gFWER control of the exceedance probabilities of  $V$  implies control of the long run relative frequency of incorrect decision across different studies.

## Partitioning to control the gFWER

One approach to controlling the gFWER is to use a method that controls the familywise error rate (i.e., controlling the gFWER at  $k = 0$ ), and then augments the rejections by automatically rejecting the null hypotheses associated with the next  $m$  extreme test statistics. This is the *augmentation* approach of van der Laan, Dudoit and Pollard (2004).

A different approach is to use the generalized partitioning principle of Xu and Hsu (2007). It partitions the parameter space into disjoint subspaces  $\Theta_I^*$  just like the partitioning principle introduced in Section 2.3.5, but adds the concept of testing *individual* hypotheses  $H_i$ . Specifically, in each  $\Theta_I^*$ , it rejects all  $H_i$ ,  $i \notin I$ , and test

$$\{H_i : \theta_i \in \Theta_i, i \in I\}$$

at gFWER level  $\alpha$ . An  $H_i$  is then rejected if it is rejected in  $\Theta_I^*$  for all  $I \ni i$ .

Note that the original partitioning principle can be viewed as a special case of the generalized partitioning principle in that, when testing in  $\Theta_I$ , the hypotheses  $H_i$ ,  $i \in I$ , are either all accepted or rejected.

A particular application of the Generalized Partitioning Principle is to use Markov's inequality to provide a gFWER-controlling test for each  $\Theta_I^*$ . Suppose the level of each marginal test is  $\alpha$ . Then, in testing  $|I|$  true null hypotheses, Markov's inequality states

$$P(V > k) \leq \frac{|I|\alpha}{k+1}.$$

Thus, one can control gFWER at level  $\alpha$  when testing in  $\Theta_I^*$  by testing each individual hypothesis at level  $\alpha(k+1)/|I|$ . The resulting multiple test is the gFWER-controlling method of Hommel and Hoffmann (1988), which was re-discovered by Lehmann and Romano (2005).

## Conditional versus unconditional inference

FDR is an unconditional error rate. It may be tempting for investigators to report conditional error rates in practice. Suppose that an investigator tested

100,000 hypotheses, using a method that controls FDR unconditionally at 1%. Then, having rejected 1,000 hypotheses, the investigator might want to state “10 out of these 1000 discoveries are expected to be false discoveries.” Such a statement, conditional on the realized rejections, is misleading since FDR is an unconditional expectation.

Step-down procedures that control the FWER and gFWER adjust for multiplicity conservatively, but only to the extent that some subset (but not all) of the null hypotheses  $H_i$ ,  $i = 1, \dots, m$ , might be true, conditional on how many of them have been rejected. This is in contrast to a single-step procedure which typically adjusts for multiplicity under the scenario that *all* the null hypotheses are true. The critical value (threshold) used by a step-down procedure, in effect, is the one corresponding to the maximum subset hypothesis  $H_I$  that could be true, conditional on data. Such conditional tests keep the true error rate as close to the desired error rate as possible, while still guaranteeing conservatism.

A form of conditional FDR error rate reporting, discussed in Efron (2007) for example, is to report an estimate of  $E(V)/r$  where  $r$  is the realized number of rejections,  $R = r$ . Note that Efron’s method estimates  $E(V)$  unconditionally.

## Taking dependence into account

Methods based on the Bonferroni inequality for FWER control (e.g., Holm’s method), or Markov’s inequality for gFWER control (e.g., the method in Lehmann and Romano, 2005), do not take joint distribution of the test statistics into account. They are generally conservative.

Some methods, such as Hochberg’s step-up method for FWER control, and Benjamini and Hochberg’s (1995) step-up method for FDR control, set critical values based on the assumption that the test statistics are independent. They are conservative under certain positive dependence structures.

For FWER control, if the test statistics have a multivariate normal or a multivariate  $t$  distributions under an intersection/partitioning hypothesis  $H_I$ , and the correlation structure has exactly or approximately a one-factor structure, then the *factor analytic* technique of Hsu (1992) is applicable. This technique amounts to modeling dependence by a latent variable. Alternatively, the variance-reduced Monte Carlo technique of Genz and Bretz (1999) can be applied. If the joint distribution of the test statistics is not multivariate normal or multivariate  $t$ , then resampling techniques can be used to compute the thresholds.

For gFWER control, assuming the test statistics have an exchangeable distribution under an intersection/partitioning hypothesis, Xu and Hsu (2007) constructed step-down methods that control gFWER while taking dependence among the test statistics into account. In analogy to closed/partitioning procedures based on the maximum test statistic, they proposed using an order statistic to test each intersection/partitioning hypothesis  $H_I$ . The technique

for computing threshold proposed by Xu and Hsu (2007) is a special case of the factor-analytic technique, as equal correlation can clearly be generated by a single latent variable. Romano and Wolfe (2007) also proposed gFWER-controlling methods based on order statistics, using a resampling technique to compute the thresholds.

For FDR reporting when the number of tests is large, Efron (2007) first transforms the test statistics so that those corresponding to the true null hypotheses become standard normal random variables. Then, to estimate  $E(V)$ , assuming pairs of transformed test statistics are bivariate normal, he extracts a one factor structure from the *multinomial* distribution of binned counts from ordered transformed test statistics.

In summary, general strategies for taking dependence into account are

- Model dependence as arising from a latent variable.
- Estimate dependence by resampling.

### 7.4.2 Multiplicity of subgroups

In searching for a subgroup of the patients for which a compound is especially efficacious, one must guard against the possibilities that the more subgroups are searched the more likely that one will “discover” such a subgroup by chance.

Let  $\mathbf{X} = (X_1, \dots, X_m)'$  represent measurements on the individual biomarkers from a typical patient. One might contemplate using a linear combination of the biomarker measurements  $\mathbf{b}'\mathbf{X} = b_1X_1 + \dots + b_mX_m$  to place patients in the  $G^+$  and  $G^-$  groups. For example,  $\mathbf{b}'\mathbf{X} > c$  puts the patient in the  $G^+$  group, while  $\mathbf{b}'\mathbf{X} \leq c$  puts the patient in the  $G^-$  group.

Let  $Y$  be an appropriate measure of efficacy of the compound. To see the danger of subgroup analysis without appropriate multiplicity adjustment, consider the model

$$Y = \beta_1X_1 + \dots + \beta_mX_m + \varepsilon,$$

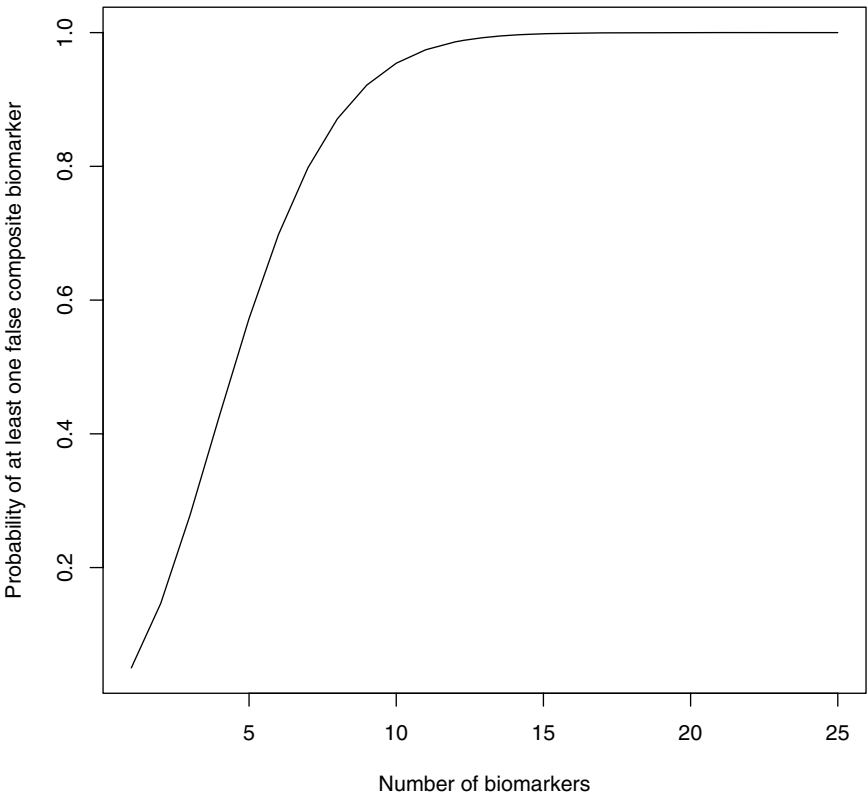
where  $\varepsilon$  represents uncertainty and assume that  $\beta_i \neq 0$  only if the  $i$ th biomarker correlates with efficacy.

Suppose none of the biomarkers correlates with efficacy (i.e.,  $\beta_i = 0$ ,  $i = 1, \dots, m$ ). Then, in testing whether individual biomarkers correlate with efficacy,  $H_i : \beta_i = 0$ , the multiplicity of having  $m$  biomarkers needs to be taken into account to avoid false positives. In testing whether any composite biomarkers in the family

$$\{H_{\mathbf{b}} : b_1\beta_1 + \dots + b_m\beta_m = 0, \mathbf{b} \in \mathfrak{R}^m\}$$

correlate with efficacy, if we assume uncertainty terms  $\varepsilon$  are independent normally distributed random errors with a known variance for simplicity, then the appropriate multiplicity adjustment to control FWER can be made by using a threshold based on a Chi-square distribution, not a normal distribution.

Figure 7.1 shows how quickly the probability of incorrectly inferring that at least one composite biomarker is correlated with efficacy approaches 100% as  $m$  increases, if each composite biomarker is tested at an error rate of  $\alpha = 0.05$ .



**FIGURE 7.1:** The probability of falsely finding a composite biomarker significant without a multiplicity adjustment.

Even if a (composite) biomarker is found to be correlated with efficacy, it does not mean that its discriminatory power would be high enough to be practically useful. Whether a compound is sufficiently efficacious for the  $G^+$  group or not requires a test on the accuracy (sensitivity or specificity) of the prediction rule based on the composite biomarker for the responder/non-responder status. In general, composite biomarkers correlated with efficacy that define subgroups can be nonlinear in the biomarker measurements. Thus, the multiplicity of subgroups can be far greater than what has been considered

in this section, and its proper adjustment would call for a different approach to the problem.

---

## 7.5 Designing pharmacogenomic studies

Microarray experiments conducted for eventual *clinical use* are, in essence, clinical trials *in silico*. As in clinical trials, designing microarray experiments according to the statistical principles below helps to ensure that the prognostic/diagnostic algorithm derived from expressions measured on the training platform applies to expressions measured on the validation platform and in its eventual clinical use.

### 7.5.1 Control for platform change using external reference sample

To control for possible systematic shifts in measured expression levels changing from the training platform to the validation platform, one can normalize expression levels between platforms using the expression levels measured from samples that are available and homogeneous from the training experiment to the validation experiment, and remain so for clinical use. An example of such samples is the Universal Reference Sample from StrataGene.

In the case of MammaPrint, the training study placed a reference sample pooled from 78 patients into one of the two channels on each of the microarrays, whereas the validation study placed a reference sample pooled from 307 patients into one of the two channels on every microarray. When the training study and validation study utilize different reference samples, it is unclear to us which is the appropriate reference sample in clinical use for individual patients.

### 7.5.2 Design to discover group differences

Measurements on gene expression levels inherently contain variability. The five sources of variability of measured gene expression levels are as follows:

1. Group: There may be differential gene expressions between risk groups (averaged over infinitely many subjects).
2. Subject: Within each group, subjects may have the same alleles but still have natural differences in expression levels (even for inbred mice).
3. Sample: Different samples from the same subject, so-called technical replicates, nevertheless will have some difference in expression levels.

4. Probe: The probes for each gene represent different parts of a gene, and will have different amounts of RNA hybridized to them.
5. Noise: Noise could come from various non-biological sources such as experimental and technical settings which may not be identical in repeated experiments.

Having replicate measurements from each level of each factor allows one to estimate each effect, and their variabilities. One can then, in turn, not only discover group differences more readily, but also estimate sensitivity and specificity of prediction algorithms, as follows.

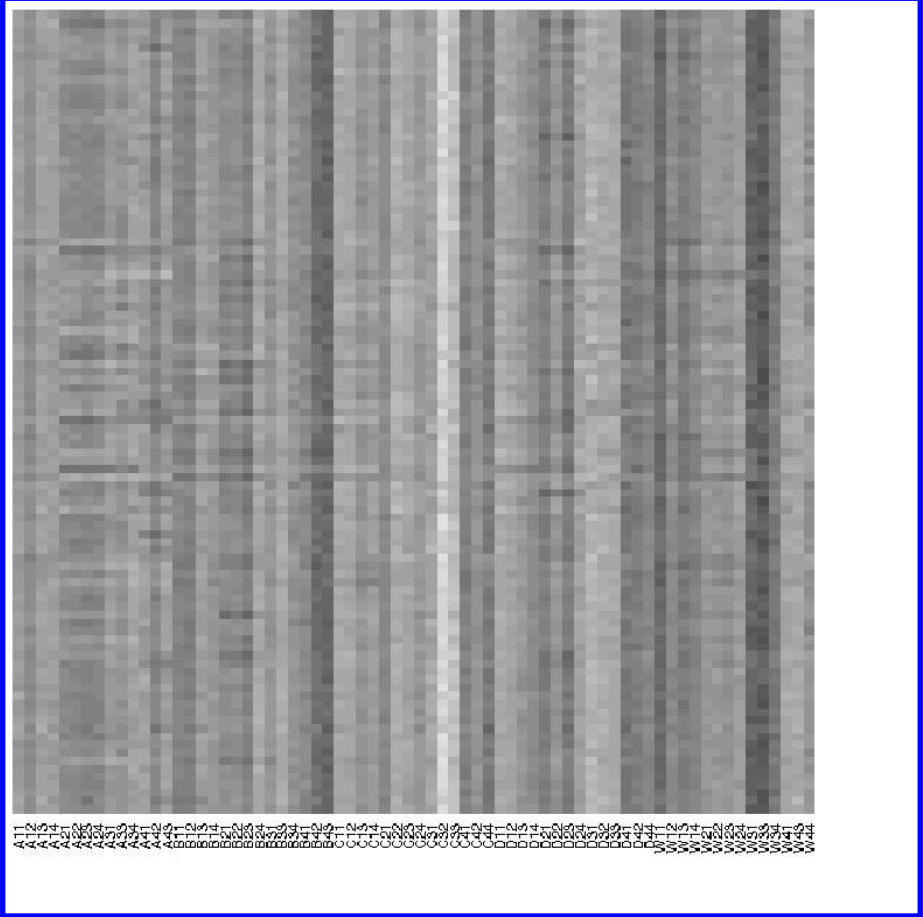
## **Replicate to estimate and remove variabilities**

Figure 7.2 displays observed expression levels of 99 genes from five groups of mice from our proof of concept experiment described in Section 7.7 (after background correction and normalization). Separation of the groups cannot be seen. The reason turns out to be that mouse and sample variabilities overwhelm group differences.

In order to estimate the effect of each subject, and its variability, replicate samples from each subject is needed. In order to estimate the effect of each sample, and its variability, each sample needs to be probed multiple times.

To discover group differences, if subject and sample effects can be estimated unbiasedly, then removing them may make group differences reveal themselves more readily. By treating subject and sample as fixed effects (as one would adjusting for covariate effects), one may more readily identify differentially expressed genes. Multiple tests conducted in this fashion control error rates conditionally, conditional on the subjects and the samples. Therefore, they control error rates unconditionally as well.

In analyzing expression levels to discover genes differentially expressed between groups, expression level is the response variable while group, subject, sample, and probe all are predictor variables. However, in training a classification algorithm based on differentially expressed genes, their expression levels then become predictors for treatment outcome. Sensitivity and specificity of such an algorithm depends on how variable expressions are between subjects within each group, and between samples within each subject. With replicate samples from each subject, and replicate probes for each sample, the variability of subject and sample can be estimated by considering them as *random* effects in modeling expression level data. Analysis of gene expressions should be cognizant of this distinction between the roles of expression levels. Figure 7.3 displays estimated expression levels of 99 genes, after estimated mouse and sample effects are removed by modeling them as fixed effects. Clustering of 80 samples with the estimated gene expression levels rediscovered the five groups of mice, arranging the columns corresponding to the 80 samples perfectly into five distinct groups.



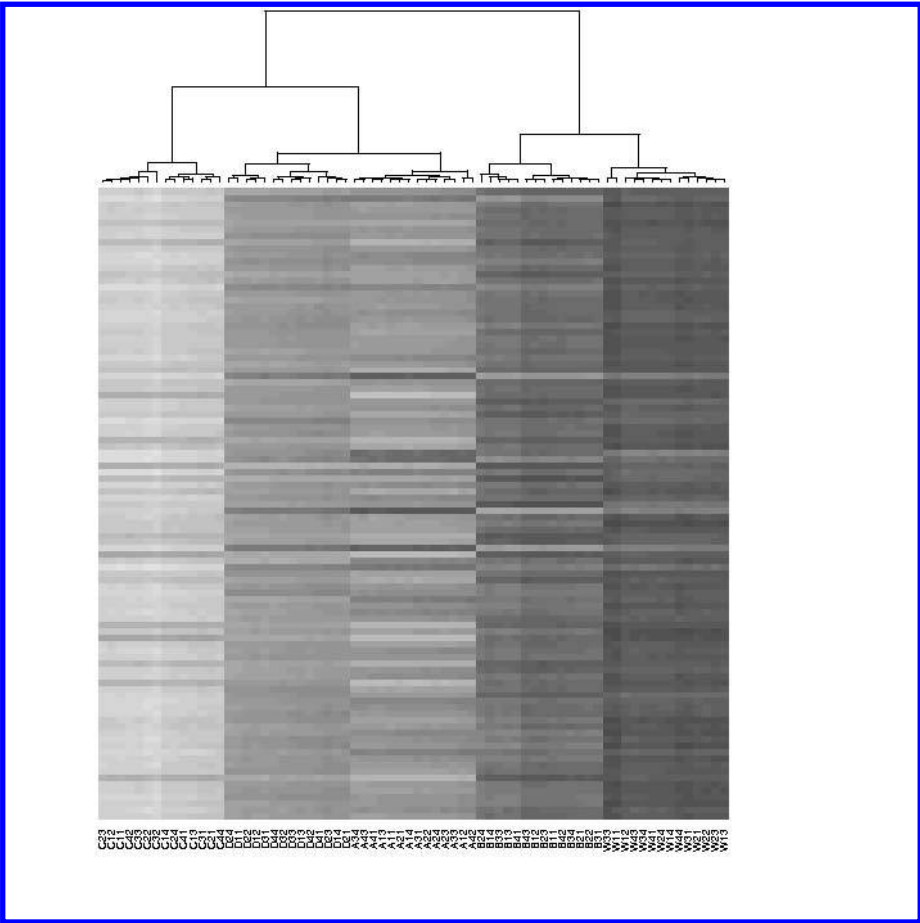
**FIGURE 7.2:** Observed expression levels of genes from inbred mice. Rows, from top to bottom, represent gene 1 to gene 99. Columns correspond to samples from five groups of mice, arranged from left to right: group A, B, C, D and W (wildtype). There are four mice per group, and four samples per mouse. The 80 columns, from left to right, represent sample 1 of mouse 1 from group A, to sample 4 of mouse 4 from group W.

### Block to avoid confounding

Gene expression measurements from microarrays are potentially affected by extraneous effects such as array or batch processing effects. Many microarray experiments take only one sample from each subject, hybridizing it to one microarray. Estimates of group differences are then potentially confounded with array effects.

In microarray experiments, a statistical *block* is a condition under which





**FIGURE 7.3:** Estimated expression levels of genes from inbred mice, with Mouse and Sample effects removed. Rows, from top to bottom, represent gene 1 to gene 99. Arranging the columns by unsupervised machine learning (clustering) results in five distinct groups: from left to right, groups C, D, A, B and W (wildtype).

measured gene expressions are likely to be equally affected by confounding factors. A block might, therefore, be a batch of arrays processed together. Some microarrays, including those made with Agilent and NimbleGen technology, can have multiple biological samples hybridized on the same array, so each array can conveniently form a block.

Statistical analysis of a block design proceeds by first comparing different risk groups within blocks, and then combining such comparisons across blocks. Such an analysis not only avoids confounding due to extraneous effects,

but also increases sensitivity and specificity because it eliminates batch/array variability in the comparisons. Basic statistical design principles also suggest keeping the proportions of samples from the groups to be compared the same across blocks. Doing so not only facilitates statistical analysis by making the effects orthogonal, but makes group comparisons more efficient as well.

### Randomize sample hybridization to avoid bias

Randomization prevents *bias*. Within each block, if placement of the samples onto microarrays or processing of the biological samples over time is not randomized, the observed differences in expression levels may contain biases due to extraneous effects. To avoid such biases, we randomize the placement of biological samples onto the microarrays and the order in which the samples are hybridized.

### Randomize probe placement to avoid change of platform issue

If placement of the probes on microarrays is not randomized, measurements from the training platform and the validation platform may have different biases. To avoid such biases in order to ensure a prediction algorithm derived from the training platform applies to expressions measured on the validation platform, we randomize the placement of probes on both platforms.

A good microarray experimental design balances the allocation of samples from the groups to be compared to wells/sub-microarrays on each microarray, and to microarrays within a batch, to avoid potential confounding array or batch effects.

### 7.5.3 Permutation tests may not control error rates

Permuting raw data across groups to be compared is often used as a tool to generate reference distributions under the null hypotheses to be tested. It is thought of as capable of taking dependence into account, as well as producing more exact inferences for small samples than methods based on asymptotics.

It turns out that, depending on application, there may be an assumption needed for permutation tests to control multiple testing error rates. At issue is, in comparing parameters of the marginal distributions of two sets of multivariate observations, the validity of permutation testing is affected by all the parameters in the joint distributions of the observations.

Calian, Li and Hsu (2008) showed the surprising fact that, in the case of a linear model with independent, identically distributed errors, permuting raw data (instead of residuals) across groups to be compared turns out to control the FWER, if the test statistic for each partitioning hypothesis is based on ordinary least squares estimates and of the maximum test statistic form.

On the other hand, as shown in Xu and Hsu (2007), in comparing the mean expression levels of genes between two groups of subjects, permuting raw

data across groups may not generate correct reference distributions under the null hypotheses, unless equalities of mean expression levels for a set of genes automatically imply equality of joint distributions of expressions levels for this set of genes (including equalities of variances, covariances, and higher cumulants).

Our view is, unless such an assumption can be made on biological ground, it is safer to take a modeling approach, an example of which is described below.

## 7.6 Analyzing microarray data by modeling

Suppose that a microarray experiment is conducted properly, adhering to the statistical design principles and addressing the issues on different sources of variability in gene expression measurements discussed so far. We can analyze the microarray data by proper modeling. A concrete example of such a design and experiment is to be given in the next section.

Let  $y_{igkspr}$  denote the background corrected, log transformed and normalized probe intensity for the  $i$ th gene ( $i = 1, \dots, n_i$ ),  $s$ th sample ( $s = 1, \dots, n_{s(k,g)}$ ) from the  $k$ th subject ( $m = 1, \dots, n_{k(g)}$ ) in group  $g$  ( $g = 1, \dots, n_g$ ),  $p$ th probe ( $p = 1, \dots, n_{p(i)}$ ), and  $r$ th replicate ( $r = 1, \dots, n_r$ ) from the experiment.

We assume, for each  $i$ ,  $y_{igkspr}$  follows a linear effects model

$$Y_{igkspr} = \mu_i + \tau_{ig} + M_{ik(g)} + S_{is(k(g))} + \pi_{p(i)} + \varepsilon_{igkspr}, \quad (7.4)$$

where

$$\begin{aligned} \mu_i &= \text{mean gene expression for the } i\text{th gene,} \\ \tau_{ig} &= \text{group } g \text{ effect on the } i\text{th gene,} \\ M_{ik(g)} &= \text{effect of subject } k \text{ in group } g \text{ on the } i\text{th gene,} \\ S_{is(k(g))} &= \text{effect of the } s\text{th sample from the } k\text{th subject in group } g \\ &\quad \text{on the } i\text{th gene,} \\ \pi_{p(i)} &= \text{effect of the } p\text{th probe on the } i\text{th gene,} \\ \varepsilon_{igkspr} &= \text{measurement error.} \end{aligned}$$

Replicate measurement errors are assumed to be independent, identically distributed with variance  $\sigma_{ie}^2$ . If they are normally distributed, then estimated group differences have a multivariate normal distribution, from which multiple tests can be derived. If they are not normally distributed, then multiple tests based on resampling of the residuals can be used.

Subject and sample effects can be considered as either fixed or random

depending on the purpose of the microarray analysis. If the subject and sample effects are random, their variances are assumed to be  $\sigma_{iM}^2$  and  $\sigma_{iS}^2$  respectively. In this case,  $\text{var}(Y_{igkspr}) = \sigma_{i\varepsilon}^2 + \sigma_{iM}^2 + \sigma_{iS}^2$ , the covariance between gene expressions for different replicates from the same sample is  $\sigma_{iM}^2 + \sigma_{iS}^2$  and the covariance between gene expressions for different samples from the same subject is  $\sigma_{iM}^2$ .

Model (7.4) is a *marginal* model in the sense that it does not specify the *joint* distribution of  $y_{igkspr}$  across the genes. Modelling expression levels from all genes simultaneously with subject by gene interaction included would generally require too much computer memory.

Other models exist for gene expression data from cDNA spotted arrays and stock Affymetrix arrays (e.g., Wolfinger et al., 2001; Chu et al., 2002, 2004; Smyth, 2004; Lee et al., 2002). However, the model (7.4) differs significantly from the existing models in that the design underlying it allows for the separate estimation of the subject and sample effects.

Consider, for example, the model Chu et al. (2004) used to analyze the ionizing radiation data in Tusher, Tibshirani and Chu (2001). Each of two samples from each of four groups (treatment and cell line combinations with two levels each) was hybridized to a stock Affymetrix array. On a gene-by-gene basis, their linear mixed effects model was

$$Y_{igpl} = \mu_i + \tau_{ig} + \pi_{p(i)} + A_{l(g)} + \varepsilon_{igpl}, \quad (7.5)$$

where  $Y_{igpl}$  is the log transformed perfect match values for probe-level data from the  $i$ th gene,  $\mu_i$  is the mean gene expression for the  $i$ th gene,  $\tau_{ig}$  is the  $g$ th group (treatment and cell line combination) effect on the  $i$ th gene,  $\pi_{p(i)}$  is the  $p$ th probe effect in the  $i$ th gene,  $A_{l(g)}$  is a random array effect, and  $\varepsilon_{igpl}$  is measurement error.

Comparing model (7.5) with model (7.4), we see the array effect  $A_l(g)$  in model (7.5) includes both the subject effect  $M_{ik(g)}$  and sample effect  $S_{is(k(g))}$  in model (7.4). That is, with stock Affymetrix arrays, sample and subject effects are completely confounded with array effects, and cannot be estimated. Therefore, treatment or group effect is confounded with array effect in this case.

Smyth (2004) applied the following simple linear *fixed* effect model to fit log-transformed intensities. For the  $i$ th gene, assume

$$E(\mathbf{Y}_i) = \mathbf{X}\boldsymbol{\alpha}_i \quad (7.6)$$

$$\text{var}(\mathbf{Y}_i) = \mathbf{W}_i\sigma_i^2, \quad (7.7)$$

where  $\mathbf{Y}_i$  is a vector containing all the transformed intensities from different samples for the  $i$ th gene,  $\mathbf{X}$  is the design matrix and  $\boldsymbol{\alpha}_i$  is a vector containing all the parameters  $\mu$ ,  $\tau$  and  $\pi$  for the  $i$ th gene.  $\mathbf{W}_i$  is assumed to be a known non-negative definite matrix. It is not entirely clear how to set values for the matrix  $\mathbf{W}_i$  in real applications.

Models (7.6) and (7.4) have different variance matrices of the vector  $\mathbf{Y}_i$ . In

particular, the variances of the elements of  $\mathbf{Y}_i$  in the model (7.6) are  $\sigma_i^2$  times the diagonals of the matrix  $W_i$ , while they are given by  $\sigma_{i\epsilon}^2 + \sigma_{iM}^2 + \sigma_{iS}^2$  (a combination of separate variance components from subject effects, sample effects, and measurement error) in the model (7.4). Also, the covariances among gene expression levels are modelled differently in (7.4) and (7.6). Model (7.4) specifies the covariance between gene expression levels of different samples from the same subject to be  $\sigma_{iM}^2$  and the covariance between gene expression levels of the same sample to be  $\sigma_{iM}^2 + \sigma_{iS}^2$ . In other words, (7.4) models the additional covariance due to measurements being from the same sample *additively*. Model (7.6), on the other hand, assumes both of the covariance between gene expression levels from the same subject and that from the same sample to be multiples of  $\sigma_i^2$ . These multiples are to be specified in the matrix  $W_i$ . However, proper specification of the multiples does not seem to be straightforward, especially when the sample and subject effects are additive.

In order to borrow information from the ensemble of genes to assist in estimating the variance of each individual gene, Smyth (2004) assumed a prior distribution on  $\sigma_i^2$ ,

$$\frac{1}{\sigma_i^2} \sim \frac{1}{d_0 s_0^2} \chi_{d_0}^2.$$

With this prior specification, (7.6) is not a gene-by-gene model any more. The unknown variance  $\sigma_i^2$  in model (7.6) can then be estimated by the posterior mean of  $\sigma_i^2$  given  $s_i^2$ , i.e.,

$$\hat{\sigma}_i^2 = \frac{d_0}{d_0 + d_i} s_0^2 + \frac{d_i}{d_0 + d_i} s_i^2.$$

The estimate  $\hat{\sigma}_i^2$  shrinks the observed variances  $s_i^2$  towards the prior values  $s_0^2$  with the extent of shrinkage determined by the relative sizes of the observed and the prior degrees of freedom  $d_i$  and  $d_0$ . This is similar in principle to Tusher, Tibshirani and Chu (2001)'s idea in SAM of modifying the estimate of  $\sigma_i$  by an offset parameter, i.e.,  $\hat{\sigma}_i = s_i + s_0$ . Smyth's offset estimate is motivated by a hierarchical model, whereas  $s_0$  in SAM is empirically chosen to be a particular percentile of all  $s_i$  values without a model or an associated distribution theory.

The two-stage ANOVA model in Lee et al. (2002) is basically the same as model (7.5) except that all effects are assumed to be fixed. They also adjust the mean square error by an offset quantity in testing for significance using  $F$ -statistics.

---

## 7.7 A proof of concept experiment

We executed a microarray experiment to prove the concept that a training experiment can be designed statistically to reliably estimate the variance

components of subject, sample, and noise separately, and that marker genes can be selected for the validation study by multiple testing with a properly controlled error rate.

To simulate the comparison of phenotype groups, our experiment compared tissues from normal mice (wild type, labeled group W) with tissues from four groups of mice (labeled groups A, B, C and D) with four different mutations of the microphthalmia transcription factor (*Mitf*) gene.

Four mice were sampled from each of the five groups. From each mouse, four cRNA samples were prepared after total RNA was isolated and biotin labeled cDNA was synthesized from a spleen tissue sample.

To prove the concept that external reference samples can be used to control for the platform change, we also prepared 16 samples of Universal Mouse Reference Sample (UMRS) from StrataGene.

To demonstrate statistical design of hybridization of samples to microarrays, we utilized NimbleGen microarrays with 12 mini-microarrays on each array. The 96 samples were hybridized to eight arrays, with the samples placed in the mini-microarrays according to the three rows by four columns patterns shown in [Figure 7.4](#), following the statistical principles of randomization, replication, and blocking.

In total, 99 genes thought to be regulated by the *Mitf* gene were selected as probes. The probe set for each gene consisted of thirty-two 24-mer probes. To demonstrate the utility of statistically designing microarrays according to the principles of randomization and replication, each probe set was replicated twice in each of the mini-microarrays, and placements of the probes were completely randomized in each mini-microarray.

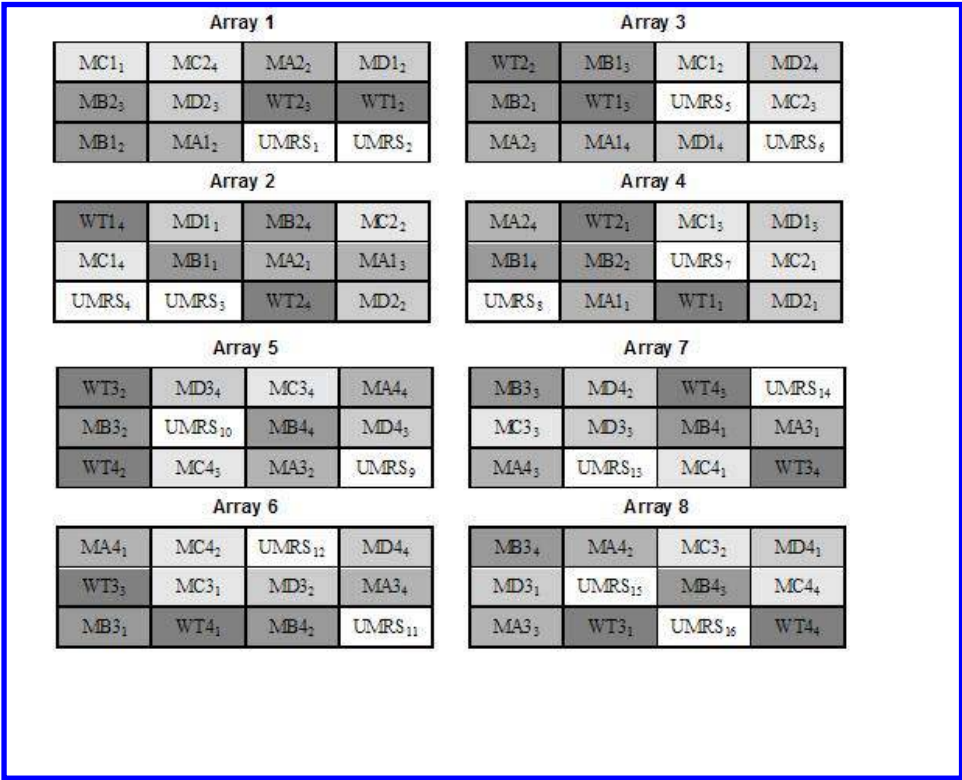
## **Normalization using an external reference sample**

Gene expression measurements from microarrays are subject to array and other processing effects, and are usually “normalized” before group comparisons are made. To “normalize” is to pre-process data to ensure observations from different sources are compatible before inferences are made.

Internal normalization uses samples within a study as controls, while external normalization uses samples external to the study as controls.

Some internal normalization techniques such as quantile normalization have been shown to be reliable within a study. However, how well inferences (such as prognostic algorithms) based on internal normalization carry across different studies is less clear.

External normalization uses reference samples that are homogeneous independent of platforms. Such external normalization can control for platform changes, provided it is as reliable within each platform as proven by internal normalization techniques. We demonstrate the viability of external normalization by showing that internal and external normalizations produce almost identical results in our study. After correcting probe level measurements for



**FIGURE 7.4:** Hybridization design in the proof of concept experiment (samples from different groups and UMRS are represented by different shades of gray, for example, MA2<sub>3</sub> is the third cRNA sample from the second mouse from group A, while UMRS<sub>2</sub> is the second UMRS).

background as described in Irizarry et al. (2003), we applied both normalization techniques, as described below.

With microarrays that allow only one biological sample to be placed on each array, it is unclear whether arrays from different groups should be normalized together or separately, due to confounding of array and group effects.

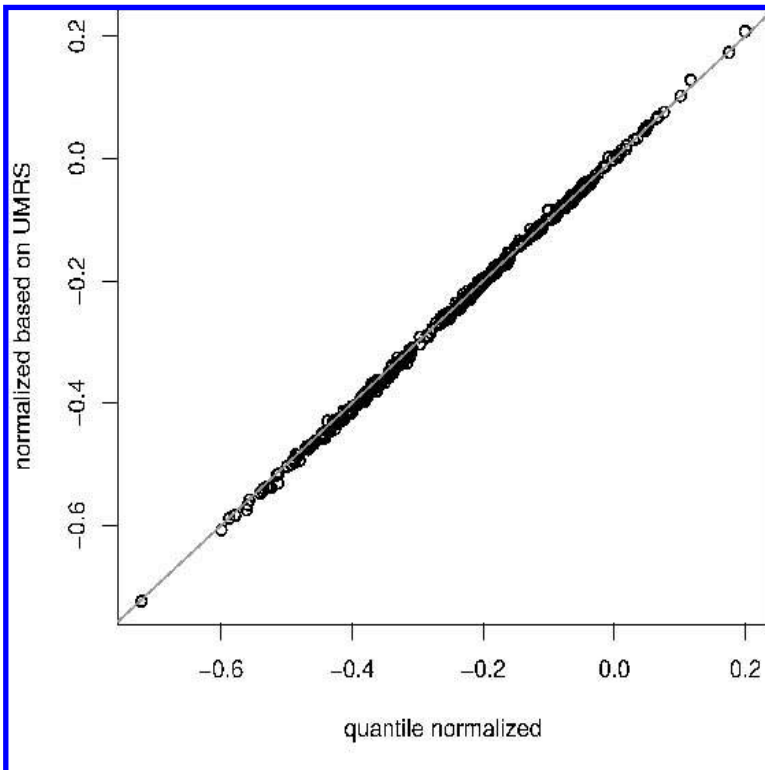
In our proof of concept experiment, however, the number of samples from each of the six groups is the same across all eight arrays. Specifically, every group of mice (groups A, B, C, D and W) and the external reference sample, UMRS, appear exactly twice on each array. It is thus reasonable to expect the distribution of the probe intensities to be the same across the arrays. We applied quantile normalization to equalize the distributions of the vectors of intensities from the eight arrays.

We propose an array-by-array external normalization process:

1. For each array, first generate a “reference” mini-microarray by averaging, for each probe, the intensities for that probe measured from UMRSs.
2. Then subtract the probe intensities in the reference mini-microarray from the corresponding probe intensities in every other (non-reference) mini-microarray.

To make an analogy to clinical trials, external normalization uses UMRS as a control.

We compared estimated differences between mutated types and wild type (A vs W, B vs W, C vs W and D vs W), after fitting the data normalized by the two techniques to the marginal model (7.4), using PROC MIXED of the SAS System. (Array 5 data was excluded due to bad quality.) Figure 7.5 shows that these two normalization techniques produce practically the same results. As quantile normalization is considered reliable, our study shows that normalization via external reference samples is a viable technique for coping with platform change issues.



**FIGURE 7.5:** Difference estimates between mutated types and wild type under quantile normalization and UMRS-based normalization.



## Multiple testing for differential expressions

To discover genes differentially expressed between each mutated group (A, B, C, D) and wildtype (W), consider testing the  $4 \times 99$  hypotheses

$$\{ \{H_{ig} : \tau_{ig} = \tau_{iW}, g = A, B, C, D\}, i = 1, 2, \dots, 99 \}. \quad (7.8)$$

in the model (7.4).

For a particular gene  $i$ , estimates of  $\tau_{ig} - \tau_{iW}, g = A, B, C, D$ , are correlated due to estimating a common  $\tau_{iW}$ . For a balanced design, if the errors are independent normally distributed, this correlation is 0.5, not negligible, and it should be taken into account in multiple testing.

For a particular group  $g$ , estimates of  $\tau_{ig} - \tau_{iW}, i = 1, \dots, 99$ , may have some correlation. Since there are  $99 \times 98/2$  correlations, the correlation structure is not easy to infer given the typical amount of data from a microarray experiment. Therefore, one can take either a conservative approach regarding these correlations or a resampling approach to adjust for the correlations.

However, for a pair of  $\tau_{ig} - \tau_{iW}$  and  $\tau_{jh} - \tau_{jW}$  involving different genes and different groups,  $i \neq j$  and  $g \neq h$ , correlation between estimates is expected to be small because the only dependence comes from the correlation between the estimates of  $\tau_{iW}$  and  $\tau_{jW}$ . Suppose this correlation equals  $\rho$ . Then, for a balanced design, the correlation between the estimates of  $\tau_{ig} - \tau_{iW}$  and  $\tau_{jh} - \tau_{jW}$  equals  $\rho/2$ .

Therefore, a practical approach to controlling the error rate in testing the  $4 \times 99$  hypotheses (7.8) is to take dependence into account in adjusting for multiplicity of testing  $\{H_{ig} : \tau_{ig} = \tau_{iW}, g = A, B, C, D\}$  for each gene  $i$  (in computing raw  $p$ -values). That is, apply the Bonferroni correction adjusting for multiplicity of the four sets of hypotheses

$$\{H_{ig} : \tau_{ig} = \tau_{iW}, i = 1, 2, \dots, 99\}, g = A, B, C, D, \quad (7.9)$$

and finally apply an appropriate method to control in testing each set of 99 hypotheses in (7.9).

Table 7.1 gives the number of genes discovered to be differentially expressed between each mutated group and wildtype, controlling gFWER at a 0.05 level, based on the linear model (7.4) with both mouse and sample effects as fixed effects. Reported are the results of applications of the Holm method, partition method using Markov's inequality, and the augmentation method for testing each set of 99 hypotheses in (7.9) at gFWER of 0.05/4 under the assumption that the errors are independent normally distributed. Alternatively, without the normal assumption, the residuals from least squares estimates were re-sampled for 10,000 times, and the results of the resampling method are also reported.

All methods find that groups  $C$  and  $W$  are most different, and groups  $B$  and  $W$  are closest in terms of the measured gene expression levels. It confirms the findings in Figure 7.3. When FWER was controlled, Holm's procedure and resampling method gave very similar results. When the number of mistakes

allowed increased to 5, resampling method came out to be much less conservative than other gFWER-controlled methods, especially for the comparison between group *B* and group *W*.

**TABLE 7.1:** Number of rejections by the step-down method controlling the gFWER at a 0.05 level, Holm method, partition method using Markov’s inequality, augmentation method, and resampling method.

<i>m</i>	Method	A vs. W	B vs. W	C vs. W	D vs. W
0	Holm	94	18	99	92
5	Markov	94	27	99	95
5	Augmentation	99	23	99	97
0	Resampling	94	17	99	92
5	Resampling	97	48	99	96

## 7.8 Software implementation

This section illustrates software implementation with SAS (linear mixed effects modeling) and R (multiple testing) for the modeling-based analyses of microarray data described in previous sections.

### 7.8.1 SAS procedures

Sensitivity and specificity of a classification algorithm depend on how variable expressions are between subjects within each group and between samples within each subject, the variability of subject and sample can be estimated by considering them as random effects in modeling expression level data. In this case, the background corrected, log transformed and normalized probe level expressions for each gene are fitted by a linear mixed effect model with subject and sample effects as random:

```
proc mixed data=bcnorm;
class group probe mice sample;
model response= group probe/DDFM=SATTERTH solution;
random mice(group) sample(group mice)/solution;
lsmeans group/diff=control('W') adjust=dunnett;
run;
```

Group effects estimates, and variance components estimates for subject effect, sample effect and measurement error for each gene are saved for sample sizes calculation in validation trial.

On the other hand, to discover group differences, if subject and sample effects can be estimated unbiasedly, then removing them may make group differences reveal themselves more readily. By treating subject and sample as fixed, one may more readily identify differentially expressed genes. In this case, the background corrected and normalized probe level expressions for each gene are fitted by a linear fixed effect model. The residuals for each gene are saved for multiple testing procedure by resampling, without assuming the errors are normally distributed.

### 7.8.2 R Functions

To discover differentially expressed genes between two groups, multiple testing by resampling the residual is applied to control the gFWER at 0.05/4. It is implemented by the following steps:

1. Resample independently with replacement the residual vectors after linear fixed effect modeling. To account for potential dependence among the measurement errors across genes, the residuals are resampled one vector at a time, with each vector consisting of residuals from within each sample. For each re-sampled data set, we compute the test statistic for each gene, which is the difference of the average intensities, averaging within each mouse and then averaging over the mice within each group. After repeated resampling  $B$  times, we have an estimated null distribution for the test statistic of each gene.
2. Calculate  $p$ -values for each gene by comparing the observed test statistics with the estimated null distribution T.mat.

```
p<-(abs(T.mat)- abs(t)>0)%*%rep(1,B)/B
```

3. Build null distribution for the  $p$ -values. Independently generate another matrix as done in step 1. The null distribution for  $p$ -values is then estimated by comparing this new matrix a.T.mat with the matrix T.mat.

```
nullP<- 1- apply(T.mat,2,fn.rawp.T, a.T.mat)
```

```
fn.rawp.T<-function(nullT, obsT)
{
  return((abs(nullT)- abs(obsT)>0)%*%rep(1,B)/B )
}
```

4. Calculate the adjusted  $p$ -value based on the generalized partitioning principal in Xu and Hsu (2007), to control the gFWER with  $m = 5$ .

```
# sort the matrix nullP so that the corresponding raw
# p-values are increasing
```

```

ind<-sort(p,index.return=T)$ix
p.H.sort<-p.H[ind, ]
P.mat.sort<-nullP[ind, ]

# step-down gFWER control
m<-5
adjp<-rep(NA,99)
# step 0
adjp[1:m]<-0
P.mat1<-P.mat.sort
# step 1
# minp here is actually p(m+1)
minp<-apply(P.mat1,2,function(x)
x[(sort(x,index.return=T)$ix)[m+1]])
adjp[m+1]<-sum(minp<p.H.sort[m+1,1])/B
# step 2 and etc.
for (i in (m+2):99)
{
P.mat1<-P.mat1[-1,]
minp<-apply(P.mat1,2, function(x)
x[(sort(x,index.return=T)$ix)[m+1]])
adjp[i]<-sum(minp<p.H.sort[i,1])/B
# enforce adjp to be increasing
adjp[i]<-max(adjp[i-1], adjp[i])
}

# present the raw p-values, adj p-values
# and the corresponding gene No.
adjp.H<-cbind(adjp, p.H.sort)
colnames(adjp.H)<-c("adjp", "rawp", "gene")
adjp.sd.gfwer<-adjp

```

## Acknowledgements

This research was supported in part by NSF Grant No. DMS-0505519 and a grant from the Icelandic Science and Technology Council.

---

## Bibliography

- [1] Abelson, R.P., Tukey, J.W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *The Annals of Mathematical Statistics*. 34, 1347-1369.
- [2] Alosch, M., Huque, M.F. (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine*. 28, 3-23.
- [3] Anderson, J.L. et al. (1999). Randomized secondary prevention trial of azithromycin in patients with coronary artery disease and serological evidence for *Chlamydia pneumoniae* infection. *Circulation*. 99, 1540-1547.
- [4] Armitage, P. (1991). Interim analysis in clinical trials. *Statistics in Medicine*. 10, 925-937.
- [5] Assmann, S.F., Pocock, S.J., Enos, L.E., Kasten, L.E. (2000). Subgroup analysis and other mis(uses) of baseline data in clinical trials. *Lancet*. 355, 1064-1069.
- [6] Bartholomew, D.J. (1959). A test of homogeneity for ordered alternatives. *Biometrika*. 46, 34-38.
- [7] Bartholomew, D.J. (1961). Ordered tests in the analysis of variance. *Biometrika*. 48, 325-332.
- [8] Bates, D.M., Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.
- [9] Bauer, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie*. 20, 130-148.
- [10] Bauer, P. (1991). Multiple testings in clinical trials. *Statistics in Medicine*. 10, 871-890.
- [11] Bauer, P. (1997). A note on multiple testing procedures in dose finding. *Biometrics*. 53, 1125-1128.
- [12] Bauer, P., Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*. 50, 1029-1041. Correction in *Biometrics*. 52, 380.

- [13] Bauer, P., König, F. (2006). Reassessment of trial perspectives from interim data—a critical view. *Statistics in Medicine*. 25, 23-36.
- [14] Bauer, P., Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine*. 18, 1833-1848.
- [15] Bauer, P., Posch, M. (2004). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections by H. Schaefer and H.-H. Mueller. *Statistics in Medicine*. 23, 1333-1334.
- [16] Bauer, P., Röhmle, J., Maurer, W., Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine*. 17, 2133-2146.
- [17] Bellamy, N. (2002). *WOMAC Osteoarthritis Index User Guide. Version V*. Brisbane, Australia.
- [18] Benjamini, Y., Hochberg Y. (1997). Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics*. 24, 407-418.
- [19] Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Ser. B*. 57, 289-300.
- [20] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (Second edition). Springer-Verlag, New York.
- [21] Berger, J. O., Wolpert, R. L. (1988). *The Likelihood Principle* (Second edition). Institute of Mathematical Statistics.
- [22] Berger, R.L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*. 24, 295-300.
- [23] Berry, D.A., Mueller, P., Grieve, A.P., Smith, M., Parke, T., Blazek, R., Mitchard, N., Krams, M. (2001). Adaptive Bayesian designs for dose-ranging drug trials. *Case Studies in Bayesian Statistics V*. Gatsonis, C., Carlin, B., Carriquiry, A. (editors). New York, Springer-Verlag, 99-181.
- [24] Bloch, D.A., Lai, T.L., Tubert-Bitter, P. (2001). One-sided tests in clinical trials with multiple endpoints. *Biometrics*. 57, 1039-1047.
- [25] Bofinger, E. (1987). Step-down procedures for comparison with a control. *Australian Journal of Statistics*. 29, 348-364.
- [26] Bolviken, E. (1982). Probability inequalities for the multivariate normal with non-negative partial correlations. *Scandinavian Journal of Statistics*. 9, 171-174.

- [27] Bornkamp, B. (2006). Comparison of model-based and model-free approaches for the analysis of dose response studies. Diploma thesis. University of Dortmund.
- [28] Bornkamp, B., Bretz, F., Dmitrienko, A., Enas G., Gaydos, B., Hsu, C.H., König, F., Krams, 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 (with discussion). *Journal of Biopharmaceutical Statistics*. 17, 965-995.
- [29] 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.
- [30] Brannath, W., Bauer, P., Posch, M. (2006). On the efficiency of adaptive designs for flexible interim decisions in clinical trials. *Journal of Statistical Planning and Inference*. 136, 1956-1961.
- [31] Brannath, W., König, F., Bauer, P. (2006). Estimation in flexible two stage designs. *Statistic in Medicine*. 25, 3366-3381.
- [32] Brannath, W., Posch, M., Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association*. 97, 236-244.
- [33] 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 in oncology. *Statistics in Medicine*. 28, 1445-1463.
- [34] Bretz, F. (2006). An extension of the Williams trend test to general unbalanced linear models. *Computational Statistics and Data Analysis*. 50, 1735-1748.
- [35] Bretz, F., Genz, A., Hothorn, L.A. (2001). On the numerical availability of multiple comparison procedures. *Biometrical Journal*. 43, 645-656.
- [36] Bretz, F., Hothorn, L.A. (2001). Testing dose-response relationships with a priori unknown, possibly non-monotonic shapes. *Journal of Biopharmaceutical Statistics*. 11, 193-207.
- [37] Bretz, F., Hothorn, L.A., Hsu, J.C. (2003). Identifying effective and/or safe doses by stepwise confidence intervals for ratios. *Statistics in Medicine*. 22, 847-858.
- [38] Bretz, F., Hothorn, T., Westfall, P. (2008). Multiple comparison procedures in linear models. *COMPSTAT 2008: Proceedings in Computational Statistics*. Brito, P. (Editor). Physica-Verlag, Heidelberg, 423-431.
- [39] Bretz, F., Hothorn, T., Westfall, P.H. (2010). *Multiple comparisons in R*. Taylor and Francis. In press.

- [40] Bretz, F., Hsu, J.C., Pinheiro, J.C., Liu, Y. (2008). Dose finding—A challenge in statistics. *Biometrical Journal*. 50, 480-504.
- [41] Bretz, F., König, F., Brannath, W., Glimm, E., Posch, M. (2009a). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*. 28, 1181-1217.
- [42] Bretz, F., Maurer, W., Brannath, W., Posch, M. (2009b). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*. 28, 586-604.
- [43] Bretz, F., Pinheiro, J., Branson, M. (2005). Combining multiple comparisons and modeling techniques in dose response studies. *Biometrics*. 61, 738-748.
- [44] Bretz, F., Schmidli, H., König, F., Racine, A., Maurer, W. (2006). Confirmatory seamless Phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*. 48, 623-634.
- [45] Buckland, S.T., Burnham, K.P., Augustin, N.H. (1997). Model selection: An integral part of inference. *Biometrics*. 53, 603-618.
- [46] Bulpitt, C.J. (1988). Subgroup analysis. *Lancet*. 2, 31-34.
- [47] Burstein, H.J. (2005). The distinct nature of HER2-positive breast cancers. *New England Journal of Medicine*. 353, 1652-1654.
- [48] Calian, V., Li, D., Hsu, J. C. (2008). Partitioning to uncover conditions for permutation tests to control multiple testing error rates. *Biometrical Journal*. 50, 756-766.
- [49] Cappizi, T., Zhang, J. (1996). Testing the hypothesis that matters for primary endpoints. *Drug Information Journal*. 30, 949-956.
- [50] Casella, G., Berger, R. L. (1990). *Statistical Inference*. Duxbury Press, New York.
- [51] Chambers, J.M., Hastie, T.J. (1992). *Statistical Models in S*. Chapman and Hall, New York.
- [52] Chevret S. (2006). *Statistical Methods for Dose Finding Experiments* (editor). Wiley, New York.
- [53] Chi, G.Y.H. (1998). Multiple testings: Multiple comparisons and multiple endpoints. *Drug Information Journal*. 32, 1347S-1362S.
- [54] Chi, G.Y.H. (2005). Some issues with composite endpoints in clinical trials. *Fundamental and Clinical Pharmacology*. 19, 609-619.



- [55] CHMP (Committee for Medicinal Products for Human Use) (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. CHMP/EWP/2459/02.
- [56] Chu, T., Weir, B., Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiment. *Mathematical Biosciences*. 176, 35-51.
- [57] Chu, T., Weir, B., Wolfinger, R. (2004). Comparison of li-wong and log-linear mixed models for the statistical analysis of oligonucleotide arrays. *Bioinformatics*. 20, 500-506.
- [58] Chuang-Stein, C., Stryszak, P., Dmitrienko, A., Offen, W. (2007). Challenge of multiple co-primary endpoints: A new approach. *Statistics in Medicine*. 26, 1181-1192.
- [59] Chuang-Stein, C., Tong, D.M. (1995). Multiple comparisons procedures for comparing several treatments with a control based on binary data. *Statistics in Medicine*. 14, 2509-2522.
- [60] Cohen, A., Sackrowitz, H.B. (1989). Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters*. 8, 273-278.
- [61] Cohen, A., Sackrowitz, H.B. (1998). Directional tests for one-sided alternatives in multivariate models. *Annals of Statistics*. 26, 2321-2338.
- [62] Cohn, J.N. et al. (1998). A dose-dependent increase in mortality with vesnarinone among patients with severe heart failure. *New England Journal of Medicine*. 339, 1810-1816.
- [63] Comelli, M., Klersy, C. (1996). Different methods to analyze clinical experiments with multiple endpoints: A comparison on real data. *Journal of Biopharmaceutical Statistics*. 6, 115-125.
- [64] Cox, D.R., Hinkley, D.V. (1974). *Theoretical Statistics*. London, Chapman and Hall.
- [65] CPMP (Committee on Proprietary Medicinal Products). (2002). Points to consider on multiplicity issues in clinical trials. CPMP/EWP/908/99.
- [66] CPMP (Committee on Proprietary Medicinal Products). (2003). Points to consider on adjustment for baseline covariates. CPMP/EWP/2863/99.
- [67] CPMP (Committee on Proprietary Medicinal Products). (2005). Guidelines on the choice of non-inferiority margin. CPMP/EWP/2158/99.
- [68] Cui, L., Hung, H.M.J., Wang, S. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*. 55, 321-324.

- [69] Cummings, S.R. et al. (1999). The effect of raloxifene on risk of breast cancer in postmenopausal women. *Journal of the American Medical Association*. 281, 2189-2197.
- [70] D'Agostino, R.B., Heeren, T.C. (1991). Multiple comparisons in over-the-counter drug clinical trials with both positive and placebo controls (with comments and rejoinder). *Statistics in Medicine*. 10, 1-31.
- [71] D'Agostino, R.B. (2000). Controlling alpha in clinical trials: The case for secondary endpoints. *Statistics in Medicine*. 19, 763-766.
- [72] Dahlöf, B. et al. (2002). Cardiovascular morbidity and mortality in the Losartan intervention for endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. *Lancet*. 359, 995-1003.
- [73] Dallow, N.S., Leonov, S.L., Roger, J.H. (2008). Practical usage of O'Brien's OLS and GLS statistics in clinical trials. *Pharmaceutical Statistics*. 7, 53-68.
- [74] Davis, C.E. (1997). Secondary endpoints can be validly analyzed, even if the primary endpoint does not provide clear statistical significance. *Controlled Clinical Trials*. 18, 557-560.
- [75] Dette, H., Bretz, F., Pepelyshev, A., Pinheiro, J.C. (2008). Optimal designs for dose finding studies. *Journal of the American Statistical Association*. 103, 1225-1237.
- [76] Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., Offen, W. (2005). *Analysis of Clinical Trials Using SAS: A Practical Guide*. SAS Press, Cary, NC.
- [77] Dmitrienko, A., Offen, W., Wang, O., Xiao, D. (2006a). Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics*. 5, 19-28.
- [78] Dmitrienko, A., Offen, W.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.
- [79] Dmitrienko, A., Tamhane, A.C. (2007). Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics*. 6, 171-180.
- [80] Dmitrienko, A., Tamhane, A.C., Liu, L. (2008). Mixtures of multiple testing procedures with gatekeeping applications. Northwestern University. Department of Industrial Engineering and Management Sciences. Working Paper 08-04.
- [81] Dmitrienko, A., Tamhane, A.C., Liu, L., Wiens, B.L. (2008). A note on tree gatekeeping procedures in clinical trials. *Statistics in Medicine*. 27, 3446-3451.

- [82] Dmitrienko, A., Tamhane, A.C., Wang, X., Chen, X. (2006b). Stepwise gatekeeping procedures in clinical trial applications. *Biometrical Journal*. 48, 984-991.
- [83] Dmitrienko, A., Tamhane, A.C., Wiens, B.L. (2008). General multistage gatekeeping procedures. *Biometrical Journal*. 50, 667-677.
- [84] Dmitrienko, A., Wiens, B.L., Tamhane, A.C., Wang, X. (2007). Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine*. 26, 2465-2478.
- [85] Dmitrienko, A., Wiens, B.L., Westfall, P.H. (2006). Fallback tests in dose-response clinical trials. *Journal of Biopharmaceutical Statistics*. 16, 745-755.
- [86] Dubey, S.D. (1985). Adjustments of  $p$ -values for multiplicities of intercorrelating symptoms. A talk presented at the Sixth International Society for Clinical Biostatisticians, Dusseldorf, Germany.
- [87] Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*. 50, 1096-1121.
- [88] 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.
- [89] Dunnett, C.W., Tamhane, A.C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association*. 87, 162-170.
- [90] Dunnett, C.W., Tamhane, A.C. (1992). Comparisons between a new drug and active and placebo controls in efficacy clinical trial. *Statistics in Medicine*. 11, 1057-1063.
- [91] Dunnett, C.W., Tamhane, A.C. (1993). Power comparisons of some step-up multiple test procedures. *Statistics and Probability Letters*. 16, 55-58.
- [92] Dunnett, C.W., Tamhane, A.C. (1995). Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics*. 51, 217-227.
- [93] Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*. 102, 93-103.
- [94] Efron, B., Morris, C. (1977). Stein's paradox in statistics. *Scientific American*. 236, 119-127.
- [95] Emerson, S. (2007). Frequentist evaluation of group sequential clinical trial designs. *Statistics in Medicine*. 26, 5047-5080.

- [96] Ettinger, B. et al. (1999). Reduction of vertebral fracture risk in postmenopausal women with osteoporosis treated with raloxifene. *Journal of the American Medical Association*. 282, 637-645.
- [97] FDA (Food and Drug Administration) (2003). *Multiplex Tests for Heritable DNA Markers, Mutations and Expression Patterns: Draft Guidance for Industry and FDA Reviewers*. Center for Devices and Radiological Health.
- [98] FDA (Food and Drug Administration) (2005a). *Drug-diagnostics co-development concept paper*. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health.
- [99] FDA (Food and Drug Administration) (2005b). *Pharmacogenomic Data Submission: Guidance for Industry*. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health.
- [100] FDA (Food and Drug Administration) (2006). *In Vitro Diagnostic Multivariate Index Assays: Draft Guidance for Industry, Clinical Laboratories, and FDA Staff*. Center for Devices and Radiological Health.
- [101] Felson, D.T., Anderson, J.J., Boers, M., Bombardier, C., Furst, D., Goldsmith, C., et al. (1995). American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. *Arthritis and Rheumatism*. 38, 727-735.
- [102] Fieller, E.C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B*. 16, 175-185.
- [103] Finner, H., Roter, M. (2001). On the false discovery rate and expected Type I errors. *Biometrical Journal*. 43, 985-1005.
- [104] Finner, H., Strassburger, K. (2002). The partitioning principle: A powerful tool in multiple decision theory. *The Annals of Statistics*. 30, 1194-1213.
- [105] Fisher, L.D. (1999). Self-designing clinical trials. *Statistics in Medicine*. 17, 1551-1562.
- [106] Fisher, L.D., Moyé, L.A. (1999). Carvedilol and the Food and Drug Administration approval process: An introduction. *Controlled Clinical Trials*. 20, 1-15.
- [107] Follman, D. (1995). Multivariate tests for multiple endpoints in clinical trials. *Statistics in Medicine*. 14, 1163-1175.
- [108] Follman, D. (1996). A simple multivariate test for one-sided endpoints. *Journal of the American Statistical Association*. 91, 854-861.

- [109] Follmann D.A., Proschan M.A., Geller N.A. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics*. 50, 325-336.
- [110] Fonseca, V., Rosenstock, J., Patwardhan, R., Salzman, A. (2002). Effect of metformin and rosiglitazone combination therapy in patients with Type 2 diabetes mellitus. *Journal of the American Medical Association*. 283, 1695-1702.
- [111] Freund, R., Littell, R. (2000). *SAS System for Regression*. Wiley, New York.
- [112] Frick, H. (1996). On the power behaviour of Läuter's exact multivariate one-sided test. *Biometrical Journal*. 38, 405-414.
- [113] Frick, H. (1997). A note on the bias of O'Brien's OLS test. *Biometrical Journal*. 39, 125-128.
- [114] Friede, T., Kieser, M. (2003). Blinded sample size reassessment in non-inferiority and equivalence trials. *Statistics in Medicine*. 22, 995-1007.
- [115] Friede, T., Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal*. 48, 537-555.
- [116] Gabriel, K.R. (1969). Simultaneous test procedures—Some theory of multiple comparisons. *Annals of Mathematical Statistics*. 40, 224-250.
- [117] Gallo, P. (2006). Operational challenges in adaptive design implementation. *Pharmaceutical Statistics*. 5, 119-124.
- [118] Geller, N.L. (2004). Design and analysis of clinical trials with multiple endpoints. *Advances in Clinical Trial Biostatistics*. Geller, N.L. (editor). Marcel Dekker, New York, 101-119.
- [119] Genz, A., Bretz, F. (1999). Numerical computation of multivariate  $t$ -probabilities with application to power calculation of multivariate contrasts. *Journal of Statistical Computation and Simulation*. 63, 361-378.
- [120] Genz, A., Bretz, F. (2002). Methods for the computation of multivariate  $t$ -probabilities. *Journal of Computational and Graphical Statistics*. 11, 950-971.
- [121] Genz, A., Bretz, F. (2009). *Computation of Multivariate Normal and  $t$  Probabilities*. Springer Verlag, Heidelberg.
- [122] Glimm, E., Srivastava, M.S., Läuter, J. (2002). Multivariate tests of normal mean vectors with restricted alternatives. *Communications in Statistics. Series B*. 31, 589-604.
- [123] Gong, J., Pinheiro, J.C., DeMets, D.L. (2000). Estimating significance level and power comparisons for testing multiple endpoints in clinical trials. *Controlled Clinical Trials*. 21, 313-329.

- [124] Gordon, A., Glazko, G., Qiu, X., Yakovlev, A. (2007). Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Annals of Applied Statistics*. 1, 179-190.
- [125] Grechanovsky, E., Hochberg, Y. (1999). Closed procedures are better and often admit a shortcut. *Journal of Statistical Planning and Inference*. 76, 79-91.
- [126] Grechanovsky, E., Pinsker, I. (1999). A general approach to step up multiple test procedures for free-combinations families. *Journal of Statistical Planning and Inference*. 82, 35-54.
- [127] Gugerli, U., Maurer, W., Mellein, B. (1993). Internally adaptive designs for parallel group trials. *Drug Information Journal*. 27, 721-732.
- [128] Guilbaud, O. (2007). Bonferroni parallel gatekeeping — transparent generalizations, adjusted  $p$ -values and short direct proofs. *Biometrical Journal*. 49, 217-227.
- [129] Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biometrical Journal*. 5, 678-692.
- [130] Hartung, H.P. et al. (2002). Mitoxantrone in progressive multiple sclerosis: a placebo-controlled, double-blind, randomised, multicentre trial. *Lancet*. 360, 2018-2025.
- [131] Hauschke, D., Pigeot, I. (2005). Rejoinder to establishing efficacy of a new experimental treatment in the gold standard design. *Biometrical Journal*. 47, 782-786.
- [132] Hayter, A.J., Hsu, J.C. (1994). On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association*. 89, 128-136.
- [133] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple significance testing. *Biometrika*. 75, 800-802.
- [134] Hochberg, Y., Mosier, M.C. (2001). Intersection-union procedures for some restricted models. Unpublished manuscript.
- [135] Hochberg, Y., Rom, D. (1995). Extensions of multiple testing procedures based on Simes' test. *Journal of Statistical Planning and Inference*. 48, 141-152.
- [136] Hochberg, Y., Tamhane, A.C. (1987). *Multiple Comparison Procedures*. Wiley, New York.
- [137] Hochberg, Y., Westfall, P.H. (2000). On some multiplicity problems and multiple comparisons procedures in biostatistics. *Handbook of Statistics*. Sen, P.K., Rao, C.R. (editors). Elsevier Sciences. 18, 75-113.

- [138] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 6, 65-70.
- [139] Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*. 75, 383-386.
- [140] Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika*. 76, 624-625.
- [141] Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal*. 43, 581-589.
- [142] Hommel, G., Bretz, F. (2008). Aesthetics and power considerations in multiple testing—A contradiction? *Biometrical Journal*. 50, 657-666.
- [143] Hommel, G., Bretz, F., Maurer, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine*. 26, 4063-4073.
- [144] Hommel, G., Hoffmann, T. (1987). Controlled uncertainty. *Multiple Hypotheses Testing*. Bauer, P., Hommel, G., Sonnemann, E. (editors). Springer, Heidelberg. 154-161.
- [145] Hommel, G., Kropf, S. (2001). Clinical trials with an adaptive choice of hypotheses. *Drug Information Journal*. 35, 1423-1429.
- [146] Hothorn, T., Bretz, F., Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*. 50, 346-363.
- [147] Hsu, J. C. (1992). The factor analytic approach to simultaneous inference in the general linear model. *Journal of Graphical and Computational Statistics*. 1, 151-168.
- [148] Hsu, J.C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics*. 12, 1136-1144.
- [149] Hsu, J.C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman and Hall, London.
- [150] 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.
- [151] Hung, H.M.J. (2000). Evaluation of a combination drug with multiple doses in unbalanced factorial design clinical trials. *Statistics in Medicine*. 19, 2079-2087.
- [152] Huque, M.F., Alosh, M. (2008). A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference*. 138, 321-335.

- [153] Huque, M.F., Alos, M. (2009). A consistency-adjusted strategy for accommodating an underpowered primary endpoint. *Biometrical Journal*. In press.
- [154] Huque, M.F., Chen, L., Sobhan, M. (2008). Statistical testing for co-primary endpoints in clinical trials. FDA/DIA Statistics Forum presentation. Bethesda, Maryland.
- [155] Huque, M.F., Sankoh, A.J. (1997). A reviewer's perspective on multiple endpoint issues in clinical trials. *Journal of Biopharmaceutical Statistics*. 7, 545-564.
- [156] ICH (International Conference on Harmonization). (1994). *Dose-response information to support drug registration (ICH E4)*.
- [157] ICH (International Conference on Harmonization). (1998). *Statistical principles for clinical trials (ICH E9)*.
- [158] Ihaka, R., Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. 5, 299-314.
- [159] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 4, 249-264.
- [160] ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. (1988). Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet*. 2, 349-360.
- [161] James, S. (1991). Approximate multinormal probabilities applied to correlated multiple endpoints in clinical trials. *Statistics in Medicine*. 10, 1123-1135.
- [162] Jennison, C., Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall, Boca Raton, FL.
- [163] Jennison, C., Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*. 22, 971-993.
- [164] Karlin, S., Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities I: Multivariate totally positive distributions. *Journal of Multivariate Analysis*. 10, 476-498.
- [165] Karlin, S., Rinott, Y. (1983).  $M$  matrices as covariance matrices of multinormal distributions. *Linear Algebra and Its Applications*. 52/53, 419-448.



- [166] Keystone, E.C. et al (2004). Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti-tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy. *Arthritis and rheumatism*. 50, 1400-1411.
- [167] Kieser, M., Bauer, P., Lehmacher, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal*. 41, 261-277.
- [168] Kieser, M., Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the Type I error rate. *Statistics in Medicine*. 22, 3571-3581.
- [169] Kieser, M., Reitmeir, P., Wassmer, G. (1995). Test procedures for clinical trials with multiple endpoints. *Biometrie in der Chemisch-in-Pharmazeutischen Industrie*. 6. Vollman, J. (editor). Fischer-Verlag, Stuttgart. 41-60.
- [170] Koch, A., Röhmle, J. (2004). Hypothesis testing in the “gold standard” design for proving the efficacy of an experimental treatment relative to placebo and a reference. *Journal of Biopharmaceutical Statistics*. 14, 315-325.
- [171] Koch, G. (1997). Discussion of “P-value adjustment for subgroup analyses”. *Journal of Biopharmaceutical Statistics*. 7, 323-331.
- [172] König, F., Brannath, W., Bretz, F., Posch, M. (2008). Adaptive Dunnett tests for treatment selection. *Statistics in Medicine*. 27, 1612-1625.
- [173] Kordzakhia, G., Dinh, P., Bai, S., Lawrence, J., Yang, P. (2008). Bonferroni-based tree-structured gatekeeping testing procedures. Unpublished manuscript.
- [174] Korn E.L., Troendle, J.F., McShane, L.M., Simon, R.S. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *Journal of Statistical Planning and Inference*. 124, 379-398.
- [175] Kotz, S., Nadarajah, S. (2004). *Multivariate t Distributions and Their Applications*. Cambridge University Press, Cambridge.
- [176] Krams, M., Lees, K.R., Hacke, W., Grieve, A.P., Orgogozo, J.M., Ford, G.A. (2003). Acute Stroke Therapy by Inhibition of Neutrophils (ASTIN). An adaptive dose-response study of UK-279,276 in acute ischemic stroke. *Stroke*. 34, 2543-2548.
- [177] Krishna R. (2006). *Dose Optimization in Drug Development*. Informa Healthcare, New York.

- [178] Kudô, A. (1963). A multivariate analog of the one-sided test. *Biometrika*. 50, 403-418.
- [179] Kwak, E.L., Clark, J.W., Chabner, B. (2007). Targeted agents: The rules of combination. *Clinical Cancer Research*. 13, 5232-5237.
- [180] Kwong, K.S., Liu, W. (2000). Calculation of critical values for Dunnett and Tamhane's step-up multiple test procedure. *Statistics and Probability Letters*. 49, 411-416.
- [181] Läuter, J. (1996). Exact  $t$  and  $F$  tests for analyzing studies with multiple endpoints. *Biometrics*. 52, 964-970.
- [182] Läuter, J., Glimm, E., Kropf, S. (1996). Exact stable multivariate tests for applications in clinical research. *Proceedings of the Biopharmaceutical Section of the American Statistical Association*. 46-55.
- [183] Lagakos, S.W. (2006). The challenge of subgroup analyses—Reporting without distorting. *New England Journal of Medicine*. 354, 1667-1669.
- [184] Lan, K.K.G., DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*. 70, 659-663.
- [185] Lang, T., Auterith, A., Bauer, P. (2000). Trend tests with adaptive scoring. *Biometrical Journal*. 42, 1007-1020.
- [186] Laska, E. M., Meisner, M. J. (1989). Testing whether an identified treatment is best. *Biometrics*. 45, 1139-1151.
- [187] Laska, E. M., Tang, D-I., Meisner, M. J. (1992). Testing hypotheses about an identified treatment when there are multiple endpoints. *Journal of the American Statistical Association*. 87, 825-832.
- [188] Lee, M. T., Lu, W., Whitmore, G., Beier, D. (2002). Models for microarray gene expression data. *Journal of Biopharmaceutical Statistics*. 12, 1-19.
- [189] Lehmacher, W., Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*. 55, 1286-1290.
- [190] Lehmacher, W., Wassmer, G., Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experiment-wise error rate. *Biometrics*. 47, 511-521.
- [191] Lehmann, E. (1986). *Testing Statistical Hypotheses* (Second edition). New York, Wiley.
- [192] Lehmann, E. L., Romano, J.P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*. 33, 1138-1154.

- [193] Li, J., Mehrotra, D.V. (2008). An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine*. 27, 5377-5391.
- [194] Li, J., Mehrotra, D. (2008). Gatekeeping testing via adaptive alpha allocation. *Biometrical Journal*. 50, 704-715.
- [195] Liu, W. (1996). Multiple tests of a non-hierarchical finite family of hypotheses. *Journal of Royal Statistical Society. Series B*. 58, 455-461.
- [196] Logan, B.R., Tamhane, A.C. (2001). Combining global and marginal tests to compare two treatments on multiple endpoints. *Biometrical Journal*. 43, 591-604.
- [197] Logan, B.R., Tamhane, A.C. (2004). On O'Brien's OLS and GLS tests for multiple endpoints. *Recent Developments in Multiple Comparison Procedures*. IMS Lecture Notes and Monograph Series. Benjamini, Y., Bretz, F., Sarkar, S. (editors), 76-88.
- [198] Logan, B.R., Tamhane, A.C. (2008). Superiority inferences on individual endpoints following non-inferiority testing in clinical trials. *Biometrical Journal*. 50, 693-703.
- [199] Müller, H.H., Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*. 57, 886-891.
- [200] Müller, H.H., Schäfer, H. (2004). A general statistical principle for changing a design anytime during the course of a trial. *Statistics in Medicine*. 23, 2497-2508.
- [201] Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., Krams, M. (2006). Adaptive seamless Phase II/III designs — Background, operational aspects, and examples. *Drug Information Journal*. 40, 463-473.
- [202] Maitournam, A., Simon, R. (2005). On the efficiency of targeted clinical trials. *Statistics in Medicine*. 24, 329-339.
- [203] Marcus, R. (1976). The power of some tests of the equality of normal means against an ordered alternative. *Biometrika*. 63, 177-183.
- [204] Marcus, R., Peritz, E., Gabriel, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*. 63, 655-660.
- [205] Maurer, W., Hothorn, L. A., Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a priori ordered hypotheses. *Biometrie in der Chemisch-in-Pharmazeutischen Industrie*. 6. Vollman, J. (editor). Fischer-Verlag, Stuttgart, 3-18.

- [206] Maurer, W., Mellein, B. (1987). On new multiple test procedures based on independent  $p$ -values and the assessment of their power. *Multiple Hypotheses Testing*. Bauer, P., Hommel, G., Sonnemann, E. (editors). Springer, Heidelberg. 48-66.
- [207] Mehrotra, D.V., Heyse, J.F. (2004). Use of false discovery rate for evaluating clinical safety data. *Statistical Methods in Medical Research*. 13, 227-238.
- [208] Morikawa, T., Yoshida, M. (1995). A useful testing strategy in Phase III trials: Combined test of superiority and test of equivalence. *Journal of Biopharmaceutical Statistics*. 5, 297-306.
- [209] Moyé, L.A. (2000). Alpha calculus in clinical trials: Considerations and commentary for the new millenium. *Statistics in Medicine*. 19, 767-779.
- [210] Moyé, L.A. (2003). *Multiple Analyses in Clinical Trials*. Springer, New York.
- [211] Moyé, L.A., Baraniuk, S. (2007). Dependence, hyper-dependence and hypothesis testing in clinical trials. *Contemporary Clinical Trials*. 28, 68-78.
- [212] Mukerjee, H., Roberston, T., Wright, F.T. (1987). Comparison of several treatments with a control using multiple contrasts. *Journal of the American Statistical Association*. 82, 902-910.
- [213] Naik, U.D. (1975). Some selection rules for comparing  $p$  processes with a standard. *Communications in Statistics. Series A*. 4, 519-535.
- [214] Neal, T. (2006). Hypothesis testing and Bayesian estimation using a sigmoid Emax model applied to sparse dose response designs. *Journal of Biopharmaceutical Statistics*. 16, 657-677.
- [215] Neuhäuser, M., Steinijans, V. W., Bretz, F. (1999). The evaluation of multiple clinical endpoints, with application to asthma. *Drug Information Journal*. 33. 471-477.
- [216] O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*. 40, 1079-1087.
- [217] Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Muirhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson J.D., Krishen, A., Liu, T., Ryder, S., Sankoh, A.J., Wang, J., Yeh, C.H. (2007). Multiple co-primary endpoints: Medical and statistical solutions. A report from the Multiple Endpoints Expert Team of the Pharmaceutical Research and Manufacturers of America. *Drug Information Journal*. 41, 31-46.

- [218] Offen, W., Helderbrand, J.D. (2003). Multiple comparison adjustments when two or more co-primary endpoints must all be statistically significant. *Proceedings of the Annual Meeting of the American Statistical Association, Biopharmaceutical Section*.
- [219] O'Neill, R.T. (1997). Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials*. 18, 550-556.
- [220] O'Neill, R.T. (2006). FDA's critical path initiative: A perspective on contributions of biostatistics. *Biometrical Journal*. 48, 559-564.
- [221] Packer, M. et al. (1996). Effect of amlodipine on morbidity and mortality in severe chronic heart failure. *New England Journal of Medicine*. 335, 1107-1114.
- [222] Packer, M., Bristow, M.R., Cohn, J.N. et al. (1996). The effects of carvedilol on morbidity and mortality in patients with chronic heart failure. *New England Journal of Medicine*. 334, 1349-1355.
- [223] Pan, G., Wolfe, D.A. (1996). Comparing groups with umbrella orderings. *Journal of the American Statistical Association*. 91, 311- 317.
- [224] Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics*. 40, 549-567.
- [225] Perlman, M.D., Wu, L. (2002). A defense of the likelihood ratio test for one-sided and order-restricted alternatives. *Journal of Statistical Planning and Inference*. 107, 173-186.
- [226] Perlman, M.D., Wu, L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics*. 60, 276-280.
- [227] Philips, J.A., Cairns, V., Koch, G.G. (1992). The analysis of a multiple-dose, combination drug clinical trial using response surface methodology. *Journal of Biopharmaceutical Statistics*. 2, 49-67.
- [228] Pigeot, I., Schäfer, J., Röhm, J., Hauschke, D. (2003). Assessing the therapeutic equivalence of two treatments in comparison with a placebo group. *Statistics in Medicine*. 22, 883-899.
- [229] Pinheiro, J., 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.
- [230] Pinheiro, J., Bretz, F., Branson, M. (2006). Analysis of dose response studies: Modeling approaches. *Dose Finding in Drug Development*. Ting, N. (editor). 146-171, Springer, New York.

- [231] Pinheiro, J.C., Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.
- [232] Pitt, B. et al. (2003). Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *New England Journal of Medicine*. 348, 1309-1321.
- [233] Pocock, S.J., Assmann, S.F., Enos, L.E., Kasten, L.E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practice and problems. *Statistics in Medicine*. 21, 2917-2930.
- [234] Pocock, S.J., Geller, N.L., Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*. 43, 487-498.
- [235] Pollard, K. S., van der Laan, M. J. (2005). Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data. *Journal of Statistical Planning and Inference*. 125, 85-100.
- [236] Posch, M., Bauer, P., Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine*. 23, 953-969.
- [237] Posch, M., Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal*. 41, 689-696.
- [238] Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Bauldauf, C., Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine*. 24, 3697-3714.
- [239] Posch, M., Timmesfeld, N. König, F., Müller, H.H. (2004) Conditional rejection probabilities of Student's *t*-test and design adaptations. *Biometrical Journal*. 46, 389-403.
- [240] Prentice, R.L. (1997). On the role and analysis of secondary outcomes in clinical trials. *Control Clinical Trials*. 18, 561-567.
- [241] Proschan, M.A., Hunsberger, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics*. 51, 1315-1324.
- [242] Proschan, M.A., Lan, K.K.G., Wittes, J.T. (2006). *Statistical Monitoring of Clinical Trials: A unified Approach*. Springer, New York.
- [243] Puri, M.L., Sen, P.K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.
- [244] Quan, H., Luo, X., Capizzi, T. (2005). Multiplicity adjustment for multiple endpoints in clinical trials with multiple doses of an active treatment. *Statistics in Medicine*. 24, 2151-2170.

- [245] Quinlan J.A., Krams M. (2006). Implementing adaptive designs: Logistical and operational considerations. *Drug Information Journal*. 40, 437-444.
- [246] Rao, Y., Lee, Y., Hsu, J. C. (2009). Determination of sample size for validation study in pharmacogenomics. Department of Statistics. The Ohio State University. In press.
- [247] Reines, S.A. et al. (2004). No effect on Alzheimer's disease in a 1-year, randomized, blinded, controlled study. *Neurology*. 62, 66-71.
- [248] Reisberg, B. et al. (2003). Memantine in moderate-to-severe Alzheimer's disease. *New England Journal of Medicine*. 348, 1333-1341.
- [249] Reitmeir, P., Wassmer, G. (1996). One-sided multiple endpoint testing in two-sample comparisons. *Communications in Statistics. Series B (Computation and Simulation)*. 25, 99-117.
- [250] Reitmeir, P., Wassmer, G. (1999). Resampling-based methods for the analysis of multiple endpoints in clinical trials. *Statistics in Medicine*. 18, 3453-3462.
- [251] Robertson, T., Wright, F.T., Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.
- [252] Rom, D. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*. 77, 663-665.
- [253] Romano, J. P., Wolf, M. (2007). Control of generalized error rates in multiple testing. *Annals of Statistics*. 35, 1378-1408.
- [254] Romano, J.P., Shaikh, A.M., Wolf, M. (2006). Formalized data snooping based on generalized error rates. Institute for Empirical Research in Economics. University of Zurich. Working Paper No. 259.
- [255] Röhmle, J., Gerlinger, C., Benda, N., Läuter, J. (2006). On testing simultaneously non-inferiority in two multiple primary endpoints and superiority in at least one of them. *Biometrical Journal*. 48, 916-933.
- [256] Röhmle, J., Pigeot, I. (2009). A comparison of multiple testing procedures for the gold standard non-inferiority trial. In press.
- [257] Roy, S.N. (1953). On a heuristic method for test construction and its use in multivariate analysis. *The Annals of Statistics*. 24, 220-238.
- [258] Royen, T. (1989). Generalized maximum range tests for pairwise comparisons of several populations. *Biometrical Journal*. 31, 905-929.
- [259] Ruberg, S.J. (1995a). Dose response studies I. Some design considerations. *Journal of Biopharmaceutical Statistics*. 5, 1-14.

- [260] Ruberg, S.J. (1995b). Dose response studies II. Analysis and interpretation. *Journal of Biopharmaceutical Statistics*. 5, 15-42.
- [261] Samuel-Cahn, E. (1996). Is the Simes' improved Bonferroni procedure conservative? *Biometrika*. 83, 928-933.
- [262] Sankoh, A.J., D'Agostino, R.B., Huque, M.F. (2003). Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues. *Statistics in Medicine*. 22, 3133-3150.
- [263] Sankoh, A.J., Huque, M.F. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*. 16, 2529-2542.
- [264] Sankoh, A.J., Huque, M.F., Russel, H.K., D'Agostino, R. (1999). Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Information Journal*. 33, 119-140.
- [265] Sarkar, S., Chang, C.K. (1997). Simes' method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*. 92, 1601-1608.
- [266] 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.
- [267] Sarkar, S.K., Snapinn, S., Wang, W. (1995). On improving the min test for the analysis of combination drug trials. *Journal of Statistical Computation and Simulation*. 51, 197-213.
- [268] Schäfer, H., Müller, H.-H. (2001). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine*. 20, 3741-3751.
- [269] Schaafsma, W., Smid, L.J. (1966). Most stringent somewhere most powerful tests against alternatives restricted by a number of linear inequalities. *The Annals of Mathematical Statistics*. 37, 1161-1172.
- [270] Schmidli, H., Bretz, F., Racine, A., Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim. Part II: Applications and Practical Considerations. *Biometrical Journal*. 48, 635-643.
- [271] Sen, P.K. (1999). Some remarks on Simes-type multiple tests of significance. *Journal of Statistical Planning and Inference*. 82, 139-145.
- [272] Senn, S. (1994). Testing for baseline imbalance in clinical trials. *Statistics in Medicine*. 13, 1715-1726.



- [273] Senn, S., Bretz, F. (2007). Power and sample size when multiple end-points are considered. *Pharmaceutical Statistics*. 6, 161-170.
- [274] Shaffer, J.P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*. 81, 826-831.
- [275] Siddiqui, M.M. (1967). A bivariate  $t$ -distribution. *The Annals of Statistics*. 38, 162-166.
- [276] Silvapulle, M.J. (1997). A curious example involving the likelihood ratio test against one-sided hypotheses. *The American Statistician*. 51, 178-180.
- [277] Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*. 63, 655-660.
- [278] Simon, R. (1982). Patient subsets and variations in therapeutic efficacy. *British Journal of Clinical Pharmacology*. 14, 473-482.
- [279] Simon, R., Maitournam, A. (2004). Evaluating the efficacy of targeted designs for randomized clinical trials. *Clinical Cancer Research*. 10, 6759-6763.
- [280] Simpson, D.G., Margolin, B.H. (1990). Nonparametric testing for dose-response curves subject to downturns: Asymptotic power considerations. *The Annals of Statistics*. 18, 373-390.
- [281] Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray. *Statistical Applications in Genetics and Molecular Biology*. 3.
- [282] Snapinn, S.M. (1987). Evaluating the efficacy of a combination therapy. *Statistics in Medicine*. 6, 657-665.
- [283] Somerville, M., Wilson, T., Koch, G., Westfall, P. (2005). Evaluation of a weighted multiple comparison procedure. *Pharmaceutical Statistics*. 4, 7-13.
- [284] Song, Y., Chi, G.Y. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine*. 26, 3535-3549.
- [285] 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). Academic Press, New York, 89-104.
- [286] Stewart, W.H., Ruberg, S.J. (2000). Detecting dose response with contrasts. *Statistics in Medicine*. 19, 913-921.

- [287] Strassburger, K., Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni based closed tests. *Statistics in Medicine*. 27, 4914-4927.
- [288] Strassburger, K., Bretz, F., Finner, H. (2007). Ordered multiple comparisons with the best and their applications to dose-response studies. *Biometrics*. 63, 1143-1151.
- [289] Strassburger, K., Bretz, F., Hochberg, Y. (2004). Compatible confidence intervals for intersection union tests involving two hypotheses. *New developments in multiple comparison procedures*. Benjamini, Y., Bretz, F., Sarkar, S. (editors). IMS Lecture Notes – Monograph Series. 47. Beachwood, Ohio, 129-142.
- [290] Tamhane, A.C., Dunnett, C.W., Green, J.W., Wetherington, J.D. (2001). Multiple test procedures for identifying the maximum safe dose. *Journal of the American Statistical Association*. 96, 835-843.
- [291] Tamhane, A.C., Dunnett, C.W., Hochberg, Y. (1996). Multiple test procedures for dose finding. *Biometrics*. 52, 21-37.
- [292] Tamhane, A.C., Liu, L. (2008). On weighted Hochberg procedures. *Biometrika*. 95, 279-294.
- [293] Tamhane, A.C., Logan B. (2002). Multiple test procedures for identifying the minimum effective and maximum safe dose of a drug. *Journal of the American Statistical Association*. 97, 293-301.
- [294] Tamhane, A.C., Logan, B. (2002). Accurate critical constants for the one-sided approximate likelihood ratio test of a normal mean vector when the covariance matrix is estimated. *Biometrics*. 58, 176-182.
- [295] Tamhane, A.C., Logan, B. (2004). A superiority-equivalence approach to one-sided tests on multiple endpoints in clinical trials. *Biometrika*. 91, 715-727.
- [296] Tamhane, A.C., Logan, B. (2006). Multiple comparison procedures for dose response studies. *Design and Analysis of Dose Response Clinical Trials*. Ting, N. (editor). Springer, New York, 172-183.
- [297] Tamhane, A.C., Logan, B. (2008). Superiority inferences on individual endpoints following non-inferiority testing in clinical trials. *Biometrical Journal*. 5, 693-703.
- [298] Tamhane, A.C., Shi, K., Strassburger, K. (2006). Power and sample size determination for a multiple test procedure for finding the maximum safe dose. *Journal of Statistical Planning and Inference*. 136, 2163-2181.

- [299] Tang, D., Geller, N.L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics*. 55, 1188-1192
- [300] Tang, D., Geller, N.L., Pocock, S.J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*. 49, 23-30.
- [301] Tang, D., Gnecco, C., Geller, N.L. (1989). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika*. 76, 577-583.
- [302] Timmesfeld, N., Schäfer, H., Müller, H.-H. (2007). Increasing the sample size during clinical trials with *t*-distributed test statistics without inflating the type I error rate. *Statistics in Medicine*. 26, 2449-2464.
- [303] Ting, N. (2006). *Dose Finding in Drug Development* (editor). Springer, New York.
- [304] Ting, N. (2003). Dose response study design. *Encyclopedia of Biopharmaceutical Statistics*. Second edition. Chow, S. (editor). Marcel Dekker, New York, 304-310.
- [305] Tsiatis, A.A., Mehta, C., (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*. 90, 367-378.
- [306] Tsiatis, A.A., Rosner, G.L., Metha, C.R. (1984). Exact confidence intervals following a group sequential test. *Biometrics*. 40, 797-803.
- [307] Tukey, J.W. (1953). *The Problem of Multiple Comparisons*. Unpublished manuscript reprinted in: *The Collected Works of John W. Tukey*. Volume 8 (1994). Braun, H.I. (editor). Chapman and Hall, New York.
- [308] 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.
- [309] Tusher, V. G., Tibshirani, R., Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*. 98, 5116-5121.
- [310] van der Laan, M.J., Dudoit, S., Pollard, K.S. (2004). Augmentation procedures for control of the generalized familywise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology*. 3, Article 15.
- [311] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 415, 530-536.

- [312] Victor, N. (1982). Exploratory data-analysis and clinical research. *Methods of Information in Medicine*. 21, 53-54.
- [313] Wakana, A., Yoshimura, I., Hamada, C. (2007). A method for therapeutic dose selection in a phase II clinical trial using contrast statistics. *Statistics in Medicine*. 26, 498-511.
- [314] Wang, R., Lagakos, S.W., Ware, J.H., Hunter, D.J., Drazen, J.M. (2007). Statistics in medicine—Reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*. 35, 2189-2194.
- [315] Wang, S.J. (1998). A closed procedure based on Follmann's test for the analysis of multiple endpoints. *Communications in Statistics, Theory and Methods*. 27, 2461-2480.
- [316] Wang, Y., McDermott, M.P. (1998). Conditional likelihood ratio test for a non-negative normal mean vector. *Journal of the American Statistical Association*. 93, 380-386.
- [317] Wassmer, G. (1999). *Statistische Testverfahren fuer gruppensequentielle und adaptive Plaene in klinischen Studien*. Verlag Alexander Moench.
- [318] Wassmer, G., Vandemeulebroecke, M. (2006). A brief review on software developments for group sequential and adaptive designs. *Biometrical Journal*. 48, 732-737.
- [319] Westfall, P. H., Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference*. 99, 25-40.
- [320] Westfall, P.H., Tobias, R.D. (2007). Multiple testing of general contrasts: Truncated closure and the extended Shaffer-Royen method. *Journal of the American Statistical Association*. 102, 487-494.
- [321] Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS Press, Cary, NC.
- [322] Westfall, P.H., Troendle, J.F. (2008). Multiple testing with minimal assumptions. *Biometrical Journal*. 50, 745-755.
- [323] Westfall, P.H., Wolfinger, R.D. (1997). Multiple tests with discrete distributions. *The American Statistician*. 51, 3-8.
- [324] Westfall, P.H., Young, S.S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley, New York.
- [325] Westfall, P.H., Zaykin, D.V., Young, S.S. (2001). Multiple tests for genetic effects in association studies. *Biostatistics Methods*. Looney, S. (editor). Humana Press, Inc.

- [326] Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Wiley, Chichester.
- [327] Wiens, B. (2003). A fixed-sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*. 2, 211-215.
- [328] Wiens, B., Dmitrienko, A. (2005). The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics*. 15, 929-942.
- [329] 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. Correction in *Biometrics*. 31, 1019.
- [330] Williams, D.A. (1972). The comparison of several dose levels with a zero dose control. *Biometrics*. 28, 519-531.
- [331] Wittes, J., Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*. 9, 65-72.
- [332] Wolfinger, R., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R. (2001). Assessing gene significance from CDNA microarray expression data via mixed models. *Journal of Computational Biology*. 8, 625-637.
- [333] Xu, H., Hsu, J.C. (2007). Applying the generalized partitioning principle to control the generalized familywise error rate. *Biometrical Journal*. 49, 52-67.
- [334] Yusuf, S., Wittes, J., Probstfield, J., Tyroler, H.A. (1991). Analysis and interpretation of subgroups of patients in randomized clinical trials. *Journal of the American Medical Association*. 266, 93-98.
- [335] Zhang, J., Quan, H., Ng, J., Stepanavage, M.E. (1997). Some statistical methods for multiple endpoints in clinical trials. *Controlled Clinical Trials*. 18, 204-221.
- [336] Zuber, E., Brannath, W., Branson, M., Bretz, F., Gallo, P., Posch, M., Racine-Poon, A. (2006). Phase II/III seamless adaptive designs with Bayesian decision tools for an efficient development of a targeted therapy in oncology. Technical report 2006-05. Department of Statistics and Decision Support Systems. University of Vienna. Available at <http://tr.smc.univie.ac.at/tr200605.html>.