Interdisciplinary Statistics

BIOEQUIVALENCE and STATISTICS in CLINICAL PHARMACOLOGY

CHAPMAN & HALL/CRC

Interdisciplinary Statistics Series

AN INVARIANT APPROACH TO STATISTICAL ANALYSIS OF SHAPES	S. Lele and J. Richtsmeier
ASTROSTATISTICS	G. Babu and E. Feigelson
BIOEQUIVALENCE AND STATISTICS IN CLINICAL PHARMACOLOGY	S. Patterson and B. Jones
CLINICAL TRIALS IN ONCOLOGY SECOND EDITION	J. Crowley, S. Green, and J. Benedetti
DESIGN AND ANALYSIS OF QUALITY OF LIFE STUDIES IN CLINICAL TRIALS	D.L. Fairclough
DYNAMICAL SEARCH	L. Pronzato, H. Wynn, and A. Zhigljavsky
GENERALIZED LATENT VARIABLE MODELING: MULTILEVEL, LONGITUDINAL, AND STRUCTURAL EQUATION MODELS	A. Skrondal and S. Rabe-Hesketh
GRAPHICAL ANALYSIS OF MULTI-RESPONSE DATA	K. Basford and J. Tukey
INTRODUCTION TO COMPUTATIONAL BIOLOGY: MAPS, SEQUENCES, AND GENOMES	M. Waterman
MARKOV CHAIN MONTE CARLO IN PRACTICE	W. Gilks, S. Richardson, and D. Spiegelhalter
MEASUREMENT ERROR AND MISCLASSIFICATION IN STATISTICS AND EPIDEMIOLOGY: IMPACTS AND BAYESIAN ADJUSTMENTS	P. Gustafson
STATISTICAL ANALYSIS OF GENE EXPRESSION MICROARRAY DATA	T. Speed
STATISTICAL CONCEPTS AND APPLICATIONS IN CLINICAL MEDICINE	J. Aitchison, J.W. Kay, and I.J. Lauder
STATISTICS FOR ENVIRONMENTAL BIOLOGY AND TOXICOLOGY	A. Bailer and W. Piegorsch
STATISTICS FOR FISSION TRACK ANALYSIS	R.F. Galbraith
STATISTICS IN MUSICOLOGY	J. Beran

Interdisciplinary Statistics

BIOEQUIVALENCE and STATISTICS in CLINICAL PHARMACOLOGY

Scott Patterson
GlaxoSmithKline Pharmaceuticals
Pennsylvania, USA

Byron Jones
Pfizer Global Research & Development
Kent, UK

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

Published in 2006 by
Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

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

International Standard Book Number-10: 1-58488-530-0 (Hardcover)
International Standard Book Number-13: 978-1-58488-530-6 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Catalog record is available from the Library of Congress

informa
Taylor & Francis Group
is the Academic Division of Informa plc.

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

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

This book is for
– my dearest Karen Rose, who has given me the most precious gift in life. (SDP)
– my family, for their support and encouragement: Hilary, Charlotte, and Alexander. (BJ)

Contents

Preface **xi**

List of figures **xiii**

List of tables **xv**

1 Drug Development and Clinical Pharmacology **1**
- 1.1 Aims of This Book 2
- 1.2 Drug Development 3
- 1.3 Clinical Pharmacology 5
- 1.4 Statistics in Clinical Pharmacology 11
- 1.5 Structure of the Book 14

2 History and Regulation of Bioequivalence **17**
- 2.1 When and How BE Studies Are Performed 19
- 2.2 Why Are BE Studies Performed? 27
- 2.3 Deciding When Formulations Are Bioequivalent 28
- 2.4 Potential Issues with TOST Bioequivalence 32
- 2.5 Current International Regulation 36

3 Testing for Average Bioequivalence **39**
- 3.1 Background 39
- 3.2 Linear Model for 2 × 2 Data 44
- 3.3 Applying the TOST Procedure 49
- 3.4 Carry-over, Sequence, and Interaction Effects 52
- 3.5 Checking Assumptions Made about the Linear Model 56
- 3.6 Power and Sample Size for ABE in the 2 × 2 Design 58
- 3.7 Example Where Test and Reference Are Not ABE 61
- 3.8 Nonparametric Analysis 68
- 3.9 Some Practical Issues 76

4 BE Studies with More Than Two Periods **79**
- 4.1 Background 80
- 4.2 Three-period Designs 81

4.3 Within-subject Variability 87
4.4 Robust Analyses for Three Period Designs 90
4.5 Four-period Designs 92
4.6 Designs with More Than Two Treatments 96
4.7 Nonparametric Analyses of Tmax 102
4.8 Technical Appendix: Efficiency 116
4.9 Tables of Data 120

5 Dealing with Unexpected BE Challenges 133
5.1 Restricted Maximum Likelihood Modelling 135
5.2 Failing BE and the DER Assessment 138
5.3 Simulation 143
5.4 Data-based Simulation 145
5.5 Carry-over 147
5.6 Optional Designs 154
5.7 Determining Trial Size 160
5.8 What Outliers Are and How to Handle Their Data 165
5.9 Bayesian BE Assessment 167
5.10 Technical Appendix 170

6 The Future and Recent Past of BE Testing 177
6.1 Brief History 178
6.2 Individual and Population BE 181
6.3 Scaled Average BE 186

7 Clinical Pharmacology Safety Studies 189
7.1 Background 191
7.2 First-time-in-humans 194
7.3 Sub-chronic Dosing Studies 208
7.4 Food-Effect Assessment and DDIs 221
7.5 Dose-Proportionality 233
7.6 Technical Appendix 240

8 QTc 243
8.1 Background 244
8.2 Modelling of QTc Data 247
8.3 Interpreting the QTc Modelling Findings 254
8.4 Design of a Thorough QTc Study in the Future 259
8.5 Technical Appendix 262

9 Clinical Pharmacology Efficacy Studies 269
9.1 Background 270
9.2 Sub-chronic Dosing 274

9.3 Phase IIa and the Proof of Concept 282
9.4 Methodology Studies 294

10 Population Pharmacokinetics 299
10.1 Population and Pharmacokinetics 300
10.2 Absolute and Relative Bioavailability 307
10.3 Age and Gender Pharmacokinetic Studies 311
10.4 Ethnicity 316
10.5 Liver Disease 321
10.6 Kidney Disease 325
10.7 Technical Appendix 328

11 Epilogue 333

Bibliography 335

Preface

This book is concerned with the use of statistics in an area of study known as clinical pharmacology. With the increasing size, duration, and cost of drug development, increased attention is being paid to this topic of clinical research with a corresponding increase in attention to the use of statistics.

This book reflects those areas of statistics which we regard as most important from a practical perspective in day-to-day clinical pharmacology work. It is not intended to be comprehensive but to provide a starting point for those engaged in research. In writing this book we have taken from our own experiences of working in the pharmaceutical industry. To emphasize this, each chapter begins with a brief vignette from Scott's experiences in the clinical pharmacology workplace. All the sets of data in the book are taken from real trials.

Following a chapter devoted to drug development and clinical pharmacology, describing the general role of statistics, we start with six chapters wholly devoted to the study of bioequivalence – a topic where successful studies are required for regulatory approval. The aim was that this should be, to a large extent, self-contained and mostly at a level that was accessible to those with some statistical background and experience.

The statistical tools developed here are useful for other topics of clinical pharmacology – namely general safety testing, testing for proarrythmic potential, population pharmacokinetics and dose selection. These topics are covered in the last four chapters of the book.

We are indebted to those who helped with their advice and discussions, over many years, on the topics of this book: Névine Zariffa, Kate Howland, Ken Waltham, Lynda Waltham, Frank Rockhold, Mike Tydeman, Darryl Downing, Lynne Collins, Dan Patterson, Vicky Patterson, Matt Huber, Andy Huber, Todd Patterson, John Whitehead, Bob Harris, Bernie Ilson, Stephen Senn, Mike Kenward, John Matthews, Dieter Hauschke, Vern Chinchilli, Frank Harrell, Lloyd Fisher, Dallas Johnson, Laszlo Endrenyi, Val Fedorov, Andy Grieve, Gary Koch, Lutz Harnisch, Vlad Dragalin, Sergei Leonov, Peter Lane, Steven Julious, Ashwini Mathur, Nick Bird, Duane Boyle, Marty Hyneck, John Finkle, Phil Sager, Delyth Jones, Paul Stober, Annabel Mortimer, Lisa Benincosa, Marty Freed, Dave Tenero, Dawn Webber, Mick Ireson, Jeff Barrett, Klaus Hinkelmann, Carl Peck, Lewis Sheiner, Nick Holford, Terry Hyslop, Walter Hauck, Marilyn Agin, Rich Anziano, Tracy Burgess, Christy Chuang-Stein, Alex Dmitrienko, Georg Ferber, Margarida Geraldes, Kalyan Ghosh, Ron Menton, Rob Muirhead, Jaya Natarajan, Walt Offen, Jay Saoud, Brian Smith, and Ram Suresh; and to those who helped us to find relevant data: Venkat Sethuraman, Ruwei Zhi, Tim Montague, Alka Preston, and Steven Kathman. We are also grateful to our employers, GlaxoSmithKline and Pfizer, for their support and permission to publish this book. Some of the plots in Chapters 3 and 4 use Splus code based on examples in [305]. SAS code to calculate the power of the test for average bioequivalence was provided by Klem Phillips. The description of the nonparametric method for estimating confidence intervals for stratified data, given at the end of Chapter 4, was based on notes written by Gunter Heimann.

This book has been typeset using the pdfLaTeX system, and we are grateful to the staff at Chapman & Hall/CRC Press for their help with this and to Rob Calver for all his assistance.

We take full responsibility for any errors or omissions in the text.

Computer software used in the text

GenStat - Sixth Edition: Lawes Agricultural Trust. Supplied by VSN International, Wilkinson House, Jordan Hill Road, Oxford, UK.

SAS: SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513, USA.

Splus 6.2 for Windows: Insightful Corporation, 1700 Westlake Avenue N, Suite 500, Seattle, Washington 98109, USA.

StatXact: Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA.

WinBUGS: MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, UK.

List of figures

1.1 The Critical Path in Drug Development 5
1.2 What Happens to the Drug After it is Taken 6
1.3 Plasma Concentration (ng/mL) versus Time (h) 7

2.1 A Normal Distribution (Mean=1, SD=1) 24
2.2 A Log-Normal Distribution 25
2.3 The Normal Distribution arising from Log-Transformation for AUC Data of Figure 2.2 26

3.1 Example 3.1: Subject Profiles Plot 42
3.2 Examples of Patterns in a Paired-agreement Plot 43
3.3 Example 3.1: Paired-agreement Plots 44
3.4 Example 3.1: Mean Differences versus Totals Plot 48
3.5 Example 3.1: Fitted Normal Densities for $\hat{\mu}_T - \hat{\mu}_R$ 52
3.6 Example 3.1: 90% Confidence Intervals for $\exp(\hat{\mu}_T - \hat{\mu}_R)$. 53
3.7 Groups-by-Periods Plot 54
3.8 Example 3.1: Groups-by-Periods Plot 57
3.9 Example 3.1: Normal Probability Plots 58
3.10 Examples of Power Curves for the TOST Procedure. Left panel: $\sigma_W = 0.1$. Right panel: $\sigma_W = 0.2$. 60
3.11 Example 3.2: Subject Profiles Plot 63
3.12 Example 3.2: Paired-agreement Plots 64
3.13 Example 3.2: Groups-by-Periods Plot 65
3.14 Example 3.2: Mean Differences versus Totals Plot 66
3.15 Example 3.2: Fitted Normal Densities for $\hat{\mu}_T - \hat{\mu}_R$ 68
3.16 Example 3.2: Normal Probability Plots 69
3.17 Example 3.1: Studentised Residuals for Tmax 73

4.1 Example 4.1: Subject Profiles Plot 84
4.2 Example 4.1: Groups-by-periods Plot 85
4.3 Example 4.1: Subject Contrasts versus Means Plot 86
4.4 Example 4.2: Subject Profiles Plot 88
4.5 Example 4.2: Groups-by-periods Plot 88
4.6 Example 4.2: Subject Contrasts versus Means Plot 90

4.7 Example 4.3: Subject Profiles Plot 94
4.8 Example 4.3: Groups-by-periods Plot 95
4.9 Example 4.3: Subject Contrasts versus Means Plot 96
4.10 Example 4.4: Subject Profiles Plot 97
4.11 Example 4.4: Groups-by-periods Plot 97
4.12 Example 4.4: Subject Contrasts versus Means Plot 99
4.13 Example 4.5: Subject Profiles Plot: Sequences 1 and 2 100
4.14 Example 4.5: Subject Profiles Plot: Sequences 3 and 4 101
4.15 Example 4.5: Subject Profiles Plot: Sequences 5 and 6 101
4.16 Example 4.5: Groups-by-periods Plot 103
4.17 Example 4.6: Subject Profiles Plot: Sequences 1 and 2 103
4.18 Example 4.6: Subject Profiles Plot: Sequences 3 and 4 104
4.19 Example 4.6: Groups-by-periods Plot 104
4.20 Example 4.5: Histograms of Bootstrap Distribution of Estimates 117

5.1 A Potentially Bioinequivalent Test and Reference Product: Fitted Normal Densities for $\hat{\mu}_T - \hat{\mu}_R$ 140

7.1 Estimated logDose versus logAUC Curve with Individual Data Points from Example 7.2.1 200
7.2 Estimated Dose versus AUC Curve (90% CI) with Individual Data Points from Example 7.2.2 203
7.3 Estimated Proportion of DLTs versus logDose from Example 7.2.3 207
7.4 Estimated Concentration versus Time (h) Profile from Phase 1 Concentration Data in `conc.sd2` 216
7.5 Dose to Exposure (AUC or Cmax) Relationship for β from 0.8 to 1.2 235

8.1 A Typical 12-Lead ECG Interval 245
8.2 Mild and Moderate QTc Prolongation ($n = 41$) in Example 8.1 250
8.3 Severe QTc Prolongation ($n = 23$) in Example 8.2 251
8.4 No QTc Prolongation ($n = 11$) in Example 8.3 252

10.1 Estimated AUCs versus Dose from a Simulated Population Pharmacokinetic Study 307
10.2 Box-plot of Estimated Clearance versus Dose from a Simulated Population Pharmacokinetic Study 308
10.3 Intrinsic and Extrinsic Population Factors Impacting the Dose-PK-PD Response Relationship [227] 317

List of tables

1.1 Regulatory Authorities 3
1.2 Potential Errors when Interpreting Bioequivalence Data 13

2.1 Schematic Plan of a 2 × 2 Cross-over Study 21
2.2 Example of a Random Allocation of Sequences to Subject in a 2 × 2 Cross-over Design 21

3.1 Example 3.1 40
3.1 Example 3.1 41
3.2 Fixed Effects in the Linear Model for the 2 × 2 Design 45
3.3 Fixed Effects: Alternative Parametrization for the 2 × 2 Design 45
3.4 Random Effects in the Linear Model for the 2 × 2 Design 46
3.5 Example 3.1: Groups-by-Periods Means (sample size in brackets) 47
3.6 Example 3.1: TOST Procedure Results 51
3.7 Fixed Effects Model Including Interactions for a 2 × 2 Design 55
3.8 Fixed Effects Model Including Interactions for a 2 × 2 Design: After Applying Constraints 55
3.9 Fixed Effects Model Including Carry-over Effects for a 2 × 2 Design 56
3.10 Fixed Effects Model Including Carry-over for a 2 × 2 Design: After Applying Constraints 56
3.11 Samples Sizes for a 2×2 Cross-over Trial to Detect ABE 60
3.12 Example 3.2 61
3.12 Example 3.2 62
3.13 Example 3.2: Groups-by-Periods Means (sample size in brackets) 66
3.14 Example 3.2: TOST Procedure Results – Log Scale 67
3.15 Example 3.2: TOST Procedure Results (all subjects) – Log Scale 67
3.16 Example 3.1: Tmax 71
3.16 Example 3.1: Tmax 72

3.17 Example 3.1: TOST Procedure Results for Tmax 73

4.1 The Expectations of $\bar{y}_{ij.}$ for Design 1 82
4.2 Example 4.1: Group-by-Period Means (sample size in brackets) 84
4.3 Example 4.1: TOST Procedure Results 87
4.4 Example 4.2: Group-by-Period Means (sample size in brackets) 89
4.5 Example 4.2: TOST Procedure Results 89
4.6 Efficiencies of Designs 1 through 7 93
4.7 Example 4.3: Group-by-Period Means (sample size in brackets) 94
4.8 Example 4.3: TOST Procedure Results 95
4.9 Example 4.4: Group-by-Period Means (sample size in brackets) 98
4.10 Example 4.4: TOST Procedure Results 98
4.11 Example 4.5: Group-by-Period Means (sample size in brackets) 102
4.12 Example 4.5: TOST Procedure Results 105
4.13 Example 4.6: Group-by-Period Means (sample size in brackets) 106
4.14 Example 4.6: TOST Procedure Results 107
4.15 Example 4.5: Tmax, Williams Design for 3 Treatments 107
4.15 Example 4.5: Tmax, Williams Design for 3 Treatments 108
4.16 Example 4.5: Tmax, Williams Design for 3 Treatments 108
4.17 Example 4.5: Tmax, Williams Design for 3 Treatments 108
4.17 Example 4.5: Tmax, Williams Design for 3 Treatments 109
4.18 Williams Design for Three Treatments: Stratified for Comparing T and R 109
4.19 Williams Design for Three Treatments: Stratified for Comparing T and S 110
4.20 Period Differences for Comparing T and R: Stratum I 113
4.21 Period Differences for Comparing T and R: Stratum II 114
4.22 Period Differences for Comparing T and R: Stratum III 115
4.23 Components of the Stratified Test for Comparing T and R 116
4.24 Efficiencies of Three Alternative Designs 120
4.25 Example 4.1: Sequence RTT 120
4.25 Example 4.1: Sequence RTT 121
4.25 Example 4.1: Sequence RTT 122
4.26 Example 4.1: Sequence TRR 122
4.26 Example 4.1: Sequence TRR 123
4.27 Example 4.2: Sequence RTT 123

4.27 Example 4.2: Sequence RTT 124
4.28 Example 4.2: Sequence TRR 125
4.29 Example 4.3 126
4.30 Example 4.4 126
4.30 Example 4.4 127
4.31 Example 4.4 127
4.32 Example 4.5: Williams Design for 3 Treatments 127
4.32 Example 4.5: Williams Design for 3 Treatments 128
4.33 Example 4.5: Williams Design for 3 Treatments 128
4.33 Example 4.5: Williams Design for 3 Treatments 129
4.34 Example 4.5: Williams Design for 3 Treatments 129
4.34 Example 4.5: Williams Design for 3 Treatments 130
4.35 Example 4.6: Williams Design for 4 Treatments 130
4.36 Example 4.6: Williams Design for 4 Treatments 131
4.37 Example 4.6: Williams Design for 4 Treatments 131
4.38 Example 4.6: Williams Design for 4 Treatments 131
4.38 Example 4.6: Williams Design for 4 Treatments 132
4.39 Example 4.6: Tmax, Williams Design for 4 Treatments 132

5.1 REML Results from PROC MIXED for Standard Bioequivalence Designs 138
5.2 REML Results from PROC MIXED for Chapter 5 Examples of Replicate Designs 139
5.3 Number of Successful BE Trials 147
5.4 Example 5.1: AUC and Cmax Data from a Replicate Cross-over Study Design with Test and Reference Formulations and Carry-over (C) 149
5.4 Example 5.1: AUC and Cmax Data from a Replicate Cross-over Study Design with Test and Reference Formulations and Carry-over (C) 150
5.4 Example 5.1: AUC and Cmax Data from a Replicate Cross-over Study Design with Test and Reference Formulations and Carry-over (C) 151
5.4 Example 5.1: AUC and Cmax Data from a Replicate Cross-over Study Design with Test and Reference Formulations and Carry-over (C) 152
5.5 Example 5.1: Adjusted Cmax and AUC Data for Subjects 1 and 10 153
5.6 Example: Variability Estimates for Use in Designing a Bioequivalence Study 164
5.7 Statistics for δ and σ_W^2 Inverse Probabilities Given AUC and Cmax Data Observed in Example 3.1 170

6.1 IBE Derived Upper 95% Bounds for Chapter 5 Replicate Design Examples 184

7.1 Schematic Plan of a First-time-in-humans Cross-over Study 197
7.2 Example 7.2.1: AUC and Cmax Data from a Cross-over First-time-in-humans Study Design 198
7.2 Example 7.2.1: AUC and Cmax Data from a Cross-over First-time-in-humans Study Design 199
7.2 Example 7.2.1: AUC and Cmax Data from a Cross-over First-time-in-humans Study Design 200
7.3 Parameter Estimates from Example 7.2.1 201
7.4 Example 7.2.2: AUC and Cmax Data from a Cross-over First-time-in-humans Study Design 202
7.5 Parameter Estimates from Example 7.2.2 203
7.6 Example 7.2.3: Dose Limiting Toxicity Data from a First-time-in-humans Trial 204
7.6 Example 7.2.3: Dose Limiting Toxicity Data from a First-time-in-humans Trial 205
7.7 Schematic Plan of a Sub-chronic Dosing Cross-over Study 209
7.8 Example 7.3.1: AUC and Cmax Data from a Sub-chronic Dosing Cross-over Study Design 210
7.8 Example 7.3.1: AUC and Cmax Data from a Sub-chronic Dosing Cross-over Study Design 211
7.9 Parameter Estimates from Example 7.3.1 213
7.10 Stationarity of Clearance Assessment from Example 7.3.1 213
7.11 Pharmacokinetic Concentration Data from Subject 47 of `conc.sd2` following a Single Dose of 5 mg 214
7.12 Estimated PK Model Parameters from Phase 1 Concentration Data in `conc.sd2` 217
7.13 ALT Data from Subject 4 of `liver.sd2` 218
7.14 Estimated ALT Data (based on `liver.sd2`) from the Sub-chronic Dosing Study Design 220
7.15 Example 7.4.1: AUC and Cmax Data from a 2 × 2 Food Effect Cross-over Study Design 223
7.16 Example 7.4.2: AUC and Cmax Data from a 2 × 2 Drug Interaction Cross-over Study Design for Metabolic Inhibition 226
7.16 Example 7.4.2: AUC and Cmax Data from a 2 × 2 Drug Interaction Cross-over Study Design for Metabolic Inhibition 227
7.17 Example 7.4.3: AUC and Cmax Data from a Drug Interaction Cross-over Study Design for Metabolic Induction 228

7.18 Example 7.5.1: AUC and Cmax Data from a Randomized Dose-Proportionality Cross-over Study 236

7.18 Example 7.5.1: AUC and Cmax Data from a Randomized Dose-Proportionality Cross-over Study 237

7.19 Exposure Estimates from a Preclinical Species 240

7.20 Exposure Estimates for a 50 kg Human from a Preclinical Species 240

8.1 First Subject's Data in Example 8.1 248

8.2 Adjusted Mean QTc (msec) and Estimated Correlation Matrix for Regimen F of Example 8.1 253

8.3 Mean Changes (90% CI) Between Regimens Following a Single Dose in Example 8.1 ($n = 41$) 254

8.4 Mean Changes (90% CI) Between Regimens Following Seven Days of Dosing in Example 8.2 ($n = 23$) 255

8.5 Mean Changes (90% CI) Between Regimens Following Seven Days of Dosing in Example 8.3 ($n = 11$) 256

8.6 Mean Changes (90% CI) Between Test Drug and Positive Control Following a Single Dose in Example 8.1 ($n = 41$) 262

8.7 Mean Changes (90% CI) Between Severe QTc Prolonger with and without a Metabolic Inhibitor Following Seven Days of Dosing in Example 8.2 ($n = 23$) 263

8.8 Rate of 'Positive' [229] Findings in 2000 Bootstrap Data Sets for Assessment of Equation (8.1) with Acceptance Boundary $\Delta = 8$ msec 264

8.9 Rate of 'Positive' [229] Findings in 2000 Bootstrap Data Sets for Assessment of Equation (8.1) with Acceptance Boundary $\Delta = 10$ msec 265

8.10 Estimated Difference in Maximum Means (msec) Between Treatments [229] in 2000 Bootstrap Data Sets Relating to Assessment of Equation (8.1) 265

8.11 Rate of Potential Hazard Findings in 2000 Bootstrap Data Sets (Lower Bound of at Least One Post-dose Comparison Exceeds Null) 266

8.12 Mean Changes (90% CI) Between Regimens Following a Single Dose in Example 8.1 ($n = 41$) Corrected for Correlation 267

9.1 Example 9.2.1: Pharmacodynamic Biomarker Data from an Exploratory Sub-chronic Dosing Study 276

9.2 Example 9.2.2: Dose, Pharmacokinetic, and Low Density Lipoprotein Data from a Sub-chronic Dosing Study 278

9.2 Example 9.2.2: Dose, Pharmacokinetic, and Low Density Lipoprotein Data from a Sub-chronic Dosing Study 279
9.3 LDL Dose-Response (Ratio relative to Baseline LDL) with 95% Confidence Intervals in Sub-chronic Dosing Study Example 9.2.2 282
9.4 Example 9.3.1: Low Density Lipoprotein Data from a Proof-of-Concept Study 283
9.4 Example 9.3.1: Low Density Lipoprotein Data from a Proof-of-Concept Study 284
9.5 LDL Effect Size (Ratio relative to Baseline LDL) from a Bayesian Statistical Analysis of a Proof-of-Concept Study Example 9.3.1 ($n = 8$) 285
9.6 Example 9.3.2: QTc Data from One Subject in a Proof-of-Concept Study 287
9.6 Example 9.3.2: QTc Data from One Subject in a Proof-of-Concept Study 288
9.7 Example 9.3.2: Plasma Pharmacokinetic-Pharmacodynamic Data from One Subject in a Proof-of-Concept Study 288
9.7 Example 9.3.2: Plasma Pharmacokinetic-Pharmacodynamic Data from One Subject in a Proof-of-Concept Study 289
9.8 QTc Response on Placebo on Days 1 and 8 in a Proof-of-Concept Study from Modelling of Dose-QTc data in Example 9.3.2 290
9.8 QTc Response on Placebo on Days 1 and 8 in a Proof-of-Concept Study from Modelling of Dose-QTc data in Example 9.3.2 291
9.9 Example 9.4.1: Biomarker Data from a Methodology Study for Subject 102 296
9.10 Example 9.4.1: Mean Changes (Simultaneous 90% Confidence Intervals) between Regimens Following a Drug Administration in a Methodology Study Corrected for Correlation 298

10.1 Estimated AUC Parameters from `conc.sd2` 303
10.2 Selected Sparse Concentration Data from Patient Studies 304
10.3 Estimated Population PK Parameters from Sparse Population Data 305
10.4 Selected Estimates for AUC from Sparse Concentration Data Obtained in Patient Studies 306
10.5 Dose-Normalized (DN) AUC from an Absolute Bioavailability Cross-over Trial 310
10.6 AUC and Cmax Data from a Pediatric (PED) and Adult (ADT) Bioavailability Trial 312

10.7 AUC and Cmax Data from a Gender Bioavailability Trial 314
10.8 Mean AUC Findings from a Gender Bioavailability Trial 315
10.9 AUC and Cmax Data from a Population Pharmacokinetic Assessment of South Korean and Western Subjects 319
10.10 Estimated Population Parameters from Evaluation of logAUC as a Function of Ethnicity, logDose, and Weight 320
10.11 Pharmacokinetic Data from a Clinical Pharmacology Hepatic Impairment Trial 322
10.11 Pharmacokinetic Data from a Clinical Pharmacology Hepatic Impairment Trial 323
10.12 Estimated Population Parameters from Evaluation of logAUC as a Function of Group 324
10.13 Pharmacokinetic Data from a Clinical Pharmacology Renal Impairment Trial 326
10.13 Pharmacokinetic Data from a Clinical Pharmacology Renal Impairment Trial 327
10.14 Estimated Population Parameters from Evaluation of logAUC as a function of Creatinine Clearance 328

CHAPTER 1

Drug Development and Clinical Pharmacology

Introducing Drug Development

It was the depths of winter, and I drove up to Philadelphia to begin working in the clinical pharmacology unit for SmithKlineBeecham Pharmaceuticals Research and Development as a brand-new biostatistician, only four days out of school. The unit is gone now, and the name of the company has changed. The folks working in ClinPharm at my company still do the same thing though – studies to bring new drug products to market and to optimise the use of drugs which are already there.

It was pretty confusing when I walked into our offices. Fresh from school, I thought the toughest part of my day was finding a parking space in West Philadelphia, but little did I know that much more fun was soon to come. Clinicians were wandering around doing clinical things, and scientists and nurses were rushing around with findings, lab samples, and dosing schedules. In the midst of all this, subjects were showing up for their studies, and getting their physical exams and being dosed.

We (the clinical pharmacology statistics group) consisted of three people then (me, my boss, and another statistician). My boss had been there for two years, and the other statistician had joined a month or two before. We were located right alongside the clinical staff, the subjects in the trials, and the laboratory personnel. It was nice to start out as a new statistician co-located with the people whom I'd work with on studies as it gave me a very practical understanding of the implications of what 'really happens' in the clinic, and we hope to convey that experience in this book.

After a couple years, though, you will prefer an office in another building. Distance makes the heart grow fonder.

My boss showed me my desk, my computer, and handed me a data set analysed by a statistician at a contract research organisation (the data set is reproduced in Chapter 3). These contract research organisations are businesses hired by drug companies to do research and/or analyses for them (i.e., on contract).

It was a collection of times in a cross-over study (see Chapters 2 and 3). She asked that I verify their findings from a nonparametric analysis *(because nobody else could, the thought was the contract research organisation had done it wrong).*

This brought several issues to mind: To what do these times correspond? What is this for? What treatments were these subjects on? Where is the rest of the study data and the protocol? What is a cross-over study (we had studied those in school, but not like this)? What is a nonparametric analysis, and which one did they use? When is lunch?

Statistically speaking, I probably should have asked the last question first. That is the first thing you need to sort out in drug development. If I had it to do over again, I would have taken a longer break before starting work too.

More important, however, is asking how such data fit into drug development, what are we trying to do with them, and what depends on the outcome. By the end of this book, you will be able to analyse these data, design studies to generate such data, know the ins and outs of drug development, and know their implications.

1.1 Aims of This Book

The purpose of this book is to provide statisticians (and other personnel in clinical pharmacology and drug development) with the methods needed to design, analyse, and interpret bioequivalence trials; when, how, and why these studies are performed as part of drug development; and to motivate the proposed methods using real world examples. The topic is a vast one and encompasses ethics, recruitment, administration, clinical operations, and regulatory issues. Some of these aspects will have a statistical component but it must be borne in mind throughout this book that the statistical features of the design and analysis are but one aspect of the role of clinical pharmacology.

Once the foundations of clinical pharmacology drug development, regulatory applications, and the design and analysis of bioequivalence trials are established, we will move to related topics in clinical pharmacology involving the use of cross-over designs. These include (but are not limited to) safety studies in Phase I, dose-response trials, drug interaction trials, food-effect and combination trials, QTc and other pharmacodynamic equivalence trials, and dose-proportionality trials.

We have tried to maintain a practical perspective and to avoid those topics that are of largely academic interest. Throughout the book we have included examples of SAS code [368] so that the analyses we describe can be immediately implemented using the SAS statistical analysis system. In particular we have made extensive use of the `proc mixed` procedure in SAS [265].

In each chapter, we will begin with the practical utility, objectives, and real-world examples of the topic under discussion. This will be followed by statistical theory and applications to support development of the area under study. Technical theory (where extensive) will be included in technical appendices to each chapter. Each topic will include worked examples to illustrate applications of the statistical techniques, their interpretation, and to serve as problems for those situations where this book serves as the basis for course-work.

1.2 Drug Development

Drug development is the process of changing someone's mind. To clarify, industrialised nations today have (pretty much uniformly) created governmental 'watch-dog' bureaucracies to regulate the use of drugs in human beings. These groups were created in response to historical events in a variety of settings where drugs which were unsafe, ineffective, or poorly made were used in human populations. Such regulatory agencies are meant to protect public health by ensuring that marketed drug products are safe, benefit the patients taking them, and are manufactured to standards of high quality (so when one takes one pill, it is the same as the next, and the next, etc).

The regulatory agencies one will frequently hear about when working in drug development are listed in Table 1.1:

Table 1.1 *Regulatory Authorities*

Nation	Agency
Australia	Therapeutic Goods Administration (TGA)
Canada	Therapeutic Products Directorate (TPD)
European Union	European Agency for the Evaluation of Medical Products (EMEA)
China	State Drug Administration (SDA)
Japan	Ministry of Health and Welfare (MHW)
United States of America	Food and Drug Administration (FDA)

These regulatory agencies are, in general, gigantic in size and scope of their activities. They employ hundreds if not thousands of people worldwide - clinicians, physicians, nurses, epidemiologists, statisticians, and a variety of other personnel. Regulatory agencies are charged with specific roles to protect the public health. Under the assumptions that all drugs are unsafe, all drugs will not benefit the patients taking them, and all drugs cannot be manufactured to high quality standards, these folks are charged with finding the few drugs that are safe, will benefit patients, and are manufactured to high quality.

No problem right? What is usually not mentioned in the charters and laws establishing these agencies are that they are also to do this as quickly as possible (folks who are sick do not like to wait) without sacrificing safety on a shoestring budget. It is a challenging job.

Drugs are chemicals or other agents (for example complex biological molecules) that have been shown to be of some benefit to public health, can be safely administered, and can be manufactured to high quality. The job of any sponsor (or drug company) is to show regulators that these three things can be done and to get their drug to patients needing it as soon as possible thereafter. In essence, drug companies are charged with changing the regulators' minds (i.e., proving them wrong). They must show that their product is safe, effective, and made to high-quality standards.

Sponsors (e.g., drug companies) develop drugs on what has been termed the critical path [137], see Figure 1.1.

A drug is generally discovered in the context of basic science - in that it causes a biological response in vitro (in a lab setting) which is thought to have the potential to provide benefit. Following an extensive battery of in vitro, animal, and manufacturing testing, and following regulatory review, it is administered to humans (a first-time-in-humans study) in a clinic. Clinical pharmacology work begins then, and extensive human and animal testing follows to evaluate safety and medical utility in parallel with scale-up of manufacturing to provide large amounts of drug substance. If all this is successful, a data package is filed with the regulatory agency where a sponsor wishes to market the product.

Generally, from the time a drug enters the clinic to the time it is approved by regulators and ready to market, 10.4 years on average elapse [90]. The cost is also substantial with estimates ranging from 0.8 to 1.7 billion dollars being spent in research and development to bring **one** new product to market [137]. Of the molecules which clear the various hurdles to human testing, only one in ten will be approved for the marketplace, failing for reasons of lack of efficacy (benefit), lack of safety, poor manufacturing, or lack of economic benefit.

What is done over this 10-plus years and a billion dollars? In a nutshell, a drug is developed by finding a dose or set of doses which produce the desired beneficial response (like lowering blood pressure: a surrogate

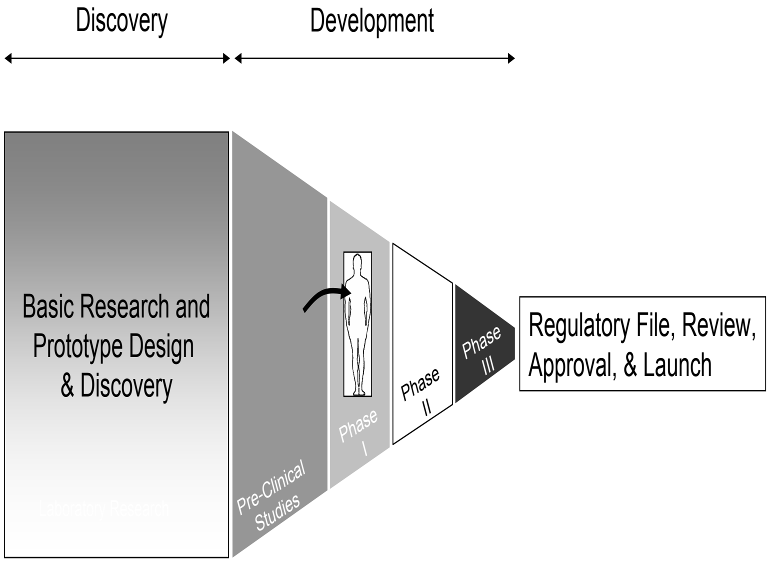
Discovery
Development
Basic Research and Prototype Design & Discovery
Laboratory Research
Pre-Clinical Studies
Phase I
Phase II
Phase III
Regulatory File, Review, Approval, & Launch

Figure 1.1 *The Critical Path in Drug Development*

marker or predictor of cardiac benefit) without producing an undesirable response (e.g., nausea, emesis). One also has to be able to make the product (manufacture it) to a standard of high quality, and in a consistent manner.

Sounds easy right? Just wait.

1.3 Clinical Pharmacology

Clinical pharmacology is the study of drugs in humans [12]. It blends the science of laboratory assessment of chemicals with the clinical and medicinal art of their application. Many textbooks are devoted to the proper study of clinical pharmacology, and we shall dwell only on those aspects which will be important for the subsequent chapters of this book.

First, some concepts. The study of pharmacokinetics (PK) is defined as 'movements of drugs within biological systems, as affected by uptake, distribution, binding, elimination and biotransformation; particularly the rates of such movements.' [413] In layman's terms, PK is what the body does to a drug (as opposed to what a drug does to the body, which we'll cover later).

When a tablet of drug is taken orally, in general, it reaches the stomach and begins to disintegrate and is absorbed (A) (Figure 1.2). When dissolved into solution in the stomach acid, the drug is passed on to the small intestine [366]. In the small intestine, many things can happen. Some of the drug will pass right on through and be eliminated (E) from the body. Some will be metabolised (M) into a different substance right in the intestine, and some drug will be picked up by the body and distributed (D) into the body through the portal circulation. This last bit of drug substance passes through the liver first, where it is often metabolized (M). The remainder passes through the liver and reaches the bloodstream where it is circulated throughout the body.

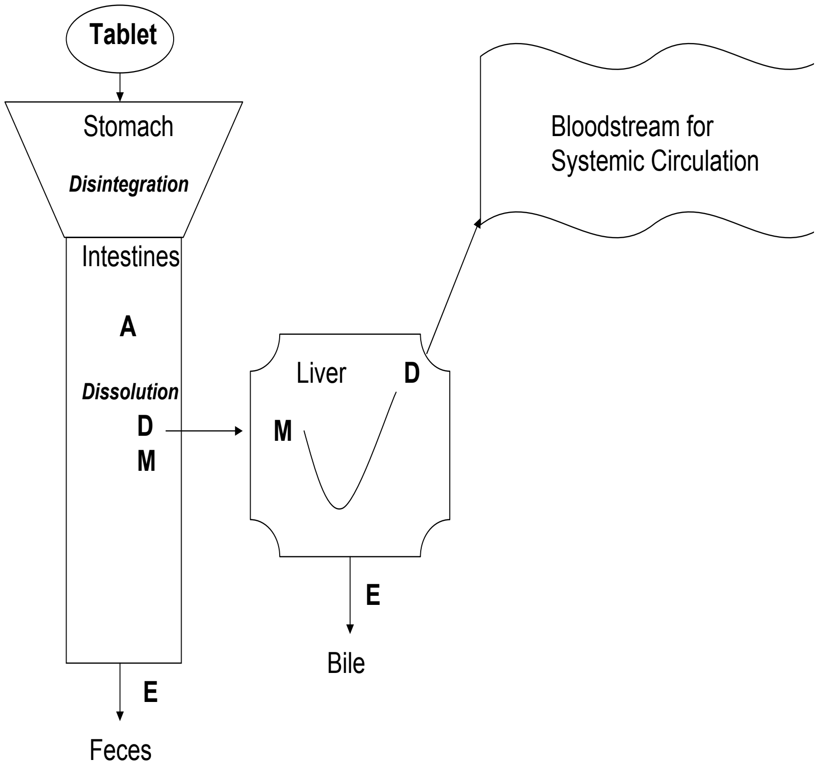

Figure 1.2 *What Happens to the Drug After it is Taken*

Following oral administration, the drug is held to undergo four 'stages' prior to being completely eliminated from the body, known as **ADME**: **A**bsorption (uptake by the body through the mouth, throat, stomach, and small/large intestine), **D**istribution (how the drug substance is carried by the body through the blood to its site of action), **M**etabolism (how the body breaks the drug substance into by-products), and **E**limination (how the body disperses the drug product). Pharmacokinetics is thus the study of ADME [12].

This process, however, is difficult to measure. Modern technology provides many options (e.g., one might tag a molecule using a radio-label and follow the progress of the molecule using X-ray imaging and similar techniques); however the most common means is to measure how much drug substance is put into the body (i.e., dose) and how much drug reaches the systemic circulation by means of blood sampling. Figure 1.3 provides a typical plasma concentration profile (vertical axis) versus time (horizontal axis) for a dose of drug given to an individual at 0 hours.

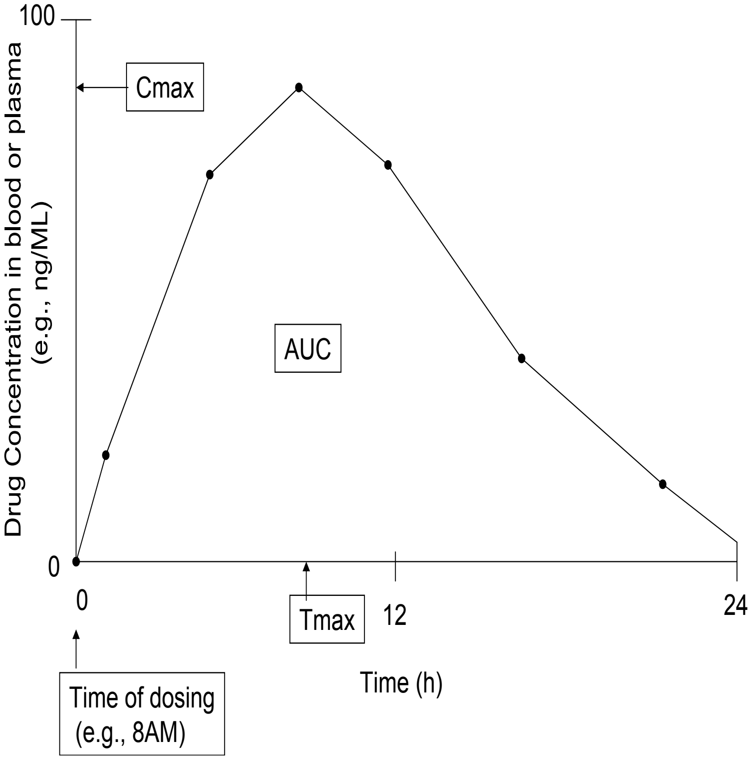

Figure 1.3 *Plasma Concentration (ng/mL) versus Time (h)*

As the drug is absorbed and distributed, the plasma concentration rises and reaches a maximum (called the Cmax or maximum concentration). Plasma levels then decline until the body completely eliminates the drug from the body. The overall exposure to drug is measured by computing the area under the plasma concentration curve (AUC). AUC is derived in general by computing the area under each time interval and adding them up.

Summary measures [366] for the plasma concentration versus time curve are derived as:

- AUC(0-t) (i.e., Area under the curve from time zero to t where t is the time of last quantifiable concentration),
- Cmax (maximum concentration),
- Tmax (time of maximum concentration),
- $T_{\frac{1}{2}}$ (half-life of drug substance), and

$$AUC(0-\infty) = AUC(0-t) + \frac{C_t}{\lambda} \tag{1.1}$$

where C_t is the concentration at time t and λ is -2.303 times the slope of the terminal phase of the $\log_e$-concentration time curve. See [417] for other summary measures. More details of techniques used in the derivation of AUC may be found in [467]. We could also fit a model to summarize a plasma concentration curve and will develop the methods used for doing so in a later chapter.

Once a drug is ingested, the substance (or active metabolite) passes through the blood and hopefully reaches a site of action; thereupon provoking what is termed a pharmacodynamic (PD) response in the body. This response is measured by looking at a biomarker or a surrogate marker.

Biomarkers are ‘a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention’, [32]. In contrast, a surrogate marker is ‘a biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit (or harm or lack of benefit or harm) based on epidemiologic, therapeutic, pathophysiologic, or scientific evidence’ [32]. Alternative definitions exist, for example, ‘a laboratory measurement or physical sign that is used in therapeutic trials as a substitute for a clinically meaningful endpoint that is a direct measure of how a patient feels, functions, or survives and is expected to predict the effect of therapy’, [425].

For example, blood pressure [425] can be considered as a surrogate marker for clinical benefit as numerous studies have shown that lowering blood pressure improves patient survival (i.e., decreases the rate of mortality seen in patients with high blood pressure). HDL (high density lipoprotein) cholesterol is a biomarker as increasing it is thought to have therapeutic cardiac benefit [411], but people are not quite entirely sure about it yet.

Large numbers of biomarkers are used in early-phase clinical development to characterize the pharmacodynamic and clinical effects of drug treatment. The purpose of clinical development at this early stage is to provide a safe and potentially effective range of doses to be fully evaluated for safety and efficacy to regulatory standards in later-phase trials (Phases IIb-IV). Generally, biomarkers are qualitatively evaluated for their predictive value in supporting later-phase development. However, recent developments highlight the need to apply quantitative tools to biomarker data to enhance their utility in support of company decisions regarding the prediction of subsequent surrogate marker and clinical outcome measures [262].

Surrogate markers have been used to support successful regulatory applications in drug development [425]. Criteria for demonstrating that an endpoint is a surrogate marker for clinical outcome are not well established [425], [262], [32]; however, some qualitative principles have been repeatedly discussed based on a publication by Temple [425]: ‘Biological Plausibility, Success in Clinical Trials, Risk-Benefit, and Public Health Considerations’.

It should be noted, however, that, at the same time as a drug is giving a ‘good’ PD response, the drug (or a metabolic by-product) may attach itself to a different site of action thereby provoking unwanted side effects. The study of pharmacodynamics is in layman’s terms ‘what the drug does to the body’.

In combination, dose, PK, and PD relationships contain the necessary and sufficient information we need to begin convincing people that use of a drug is worthwhile and to learn about the behavior of a drug product. This is sometimes also referred to as the dose-exposure-response (DER) relationship [225], [134], [399].

How do we go about developing drugs under this approach to clinical pharmacology? Early-stage development should focus on learning about the compound, understanding its safety and efficacy in patients by means of varying dose and measuring PK and PD. Once sufficient confidence is reached that the compound does what is beneficial and is safe enough to dose, sponsors begin conducting large confirmatory trials. These are trials designed to convince regulatory authorities that the drug is safe to use in the marketplace and will be of public benefit. A more comprehensive review may be found in [402].

Let us revisit our earlier discussion of drug development (Figure 1.1) in light of what we now know about clinical pharmacology and to break down the critical path of clinical drug development in more detail. Prior to the first-in-human study in clinic, in vitro and animal preclinical experimentation should establish a range of safe doses for study in humans. Doses are then selected for introduction into clinical studies in humans [356].

Clinical development of a drug product, with the exception of only the most toxic products targeted for the treatment of cancer, then initiates with the study of the drug product in normal healthy male volunteers in what is known as Phase I. These studies are typically small, well-controlled, data-intensive, dose escalating, and placebo-controlled (we will get into this in a later chapter).

In this stage of human drug development, the primary objective of a clinical study is to determine a safe range of doses and dosing regimens (e.g., once-a-day or twice-a-day) for later dosing in studies involving patients with the disease state under study. Dose and dosing regimen are examined with respect to their impact on the pharmacokinetics of the drug product. Additionally, should biomarker or surrogate markers be present to characterize the activity of the drug in normal healthy volunteers, these data are characterized relative to dose and PK.

By the end of Phase I, dose-finding studies in normal healthy volunteers or patient studies (e.g., for oncology compounds) should provide: a range of safe (and potentially efficacious) doses for further study in patients, an initial description of pharmacokinetic exposure levels and/or biomarker/surrogate marker levels at each dose to facilitate choice of dose, dose titration, dosing intervals for Phase II studies, and the development of initial models for use in pharmacokinetic-pharmacodynamic modelling for both desirable and undesirable effects.

Subsequent Phase II clinical studies in patients establish the minimum starting and maximum effective dose as well as the maximum tolerated dose in patients with the disease state using pharmacodynamic endpoints or surrogate markers of therapeutic response. Dose titration and the length of time needed to see an effect (desirable or undesirable) are also established. In these studies, models relating dose to PK and to PD are developed to understand the mechanism of the drug's action and to search for relevant covariates (e.g., age or gender) to control later Phase II or Phase III confirmatory trial designs [225].

Dose-finding studies in Phase II in the target population should establish the therapeutic window by identifying a minimum effective starting dose (the lowest dose yielding a desirable effect), a maximum effective dose (the dose beyond which further escalation lacks further desirable benefit), and a maximum tolerated dose (the dose beyond which there is an unacceptable increase in undesirable effects) in the target population. In addition, these studies should identify the time interval needed to see an effect (desirable and/or undesirable) and reasonable, responseguided, titration steps along with the time intervals at which to dose titrate, to develop updated pharmacokinetic-pharmacodynamic models for both desirable and undesirable effects in the population of interest, and to identify potential covariates to be studied for dose adjustment in Phase III (e.g., age, gender).

Once a dose or set of efficacious doses are chosen from Phase II trials and the characteristics of Figure 1.4 are mapped out, confirmatory Phase III trials are performed to support regulatory acceptance. These trials, in large numbers of patients with the disease under study, should characterize the risk relative to benefit in clinical use of the compound. These studies in Phase III should be used to establish the risk:benefit ratio and pharmacokinetic-pharmacodynamic relationship (if any) for doses chosen to be in the therapeutic window established in Phase II.

Additional clinical pharmacology studies will also be conducted in Phase III to determine how to dose the drug in patients with particular health problems (like kidney disease) and for patients taking a variety of concomitant medications. Additionally, clinical pharmacology studies will be done to confirm that new formulations of drug product are equivalent to those used in clinical development when scale-up of the manufacturing process for mass production occurs. These are bioequivalence studies and will be the subject of Chapters 2-6.

1.4 Statistics in Clinical Pharmacology

What is a statistic? It is numerical information about a given object or event. This information is derived from a sample (a study or trial) of a population (as it would often be impossible to collect information from an entire large population, that is too numerous for exhaustive measurement). On its own, a statistic is just a number. However, decisions are made based on statistics, and that is where the statistician's skill and art come into play.

James Bernoulli described nine 'general rules dictated by common sense' [175] (see Chapter 15 on Bernoulli's Ars Conjectandi, 1713) for making decisions based on statistics, and most statisticians follow these (in principle):

1. One must not use conjecture (i.e., use statistics) in cases where complete certainty is obtainable.
2. One must search for all possible arguments or evidence concerning the case (i.e., show due diligence).
3. One must take into account both arguments for and against the case.
4. For a judgment about general events, general arguments are sufficient; for individual events, however, special and individual arguments have to be taken into account.

5. In case of uncertainty, action should be suspended until more information is at hand; however, if circumstances permit no delay, the action that is most suitable, safe, wise, and probable should be chosen.
6. That which can be useful on some occasion and harmful on no occasion is to be preferred to that which is useful and harmful on no occasion.
7. The value of human actions must not be judged by their outcome.
8. In our judgments we must be wary of attributing more weight to a thing than its due and of considering something that is more probable than another to be absolutely certain.
9. Absolute certainty occurs only when the probability nearly equals the whole certainty (i.e., when the probability of some event is equal to one, such that we know it will occur).

Statisticians are applied mathematicians. In drug development, these people are responsible for quantifying the uncertainty inherent in the scientific and regulatory process of developing new drug products. The focus of our discussion will be on the techniques statisticians apply to the design and analysis of clinical pharmacology trials, but statisticians are involved in a variety of other topics associated with drug development (see [151] for more details.)

As any statistic is derived from a sample, there is **always** uncertainty involved in its use. There is always a chance that the sample and the statistic derived from it got something 'wrong' relative to the truth of the situation. Statisticians and the art of statistics are therefore employed in drug development to ensure that the probability of a 'wrong-answer' is quantified and understood so that the implications can be considered.

Consider the main topic of this book, bioequivalence. At certain times in drug development, drug companies must show that a new formulation of drug (i.e., a new capsule or tablet) is equivalent to an old formulation. It is assumed (i.e., the hypothesis) that the formulations are not equivalent, and a study must be performed to generate data to show that they are. Chapters 2, 3, and 4 will go into more detail.

Obviously it is completely impossible to assess every new tablet and compare each one to each and every old tablet to ensure high quality is present. It would take forever and be too time consuming to even contemplate, and even if we could devise a test to ensure that each and every tablet is exactly the same as each and every old one, we are more interested in if the two formulations will give us the same results when patients take them anyway. So it may not matter if they are not exactly the same.

Therefore, a clinical study is used to do the job. Data are generated in the study, and statistics are derived to compare the results of the new

		The Truth	
		Formulations are NOT equivalent	Formulations ARE equivalent
Statistics from study	Formulations are NOT equivalent	Right answer!	Wrong answer (Type 2 error)
show that	Formulations ARE equivalent	Wrong answer (Type 1 error)	Right answer!

Table 1.2 *Potential Errors when Interpreting Bioequivalence Data*

formulation to the old formulation. When the data come in, we use them to decide if we have sufficient evidence to throw out our hypothesis (that the formulations are not equivalent) and that we have sufficient data to conclude they are.

We approach this topic like a regulator would - i.e., assume that they are not equivalent until data shows that they are. The two formulations may in fact (i.e., in truth) be equivalent, but until we have conclusive data to show that, it is best to err on the side of caution.

When the data come in, they will give us information to conclude whether the drugs are equivalent or not. We can make two errors in this situation (see Table 1.2). We can conclude from the data that they are equivalent, when in fact they are not (a Type 1 error), or we could conclude that the formulations are not equivalent when in fact they are (a Type 2 error). Bernoulli's second and third principles are applied in this manner, and we will get into the application of the other Bernoulli principles in this setting later in the book.

Statisticians use tools to design and analyse studies to ensure that the probabilities of a Type 1 or 2 error are controlled and held at a quantified rate of occurrence. These tools are randomisation, replication, blocking, blinding, and modelling, and their definition and specific application will be discussed in great detail in later chapters. Application of these tools enables those using the statistics (i.e., the drug companies and regulators) to know the implications of their decision on whether the two formulations are equivalent or not and to make a reasoned decision on whether to provide the new formulation to the patients who need it.

Not much is 100% certain, and studies like those described above are no exception. It is not unusual for studies to give misleading (i.e., Type

1 or 2 error) results when one considers that thousands of clinical trials are performed worldwide each year. Even as small an error rate as 5% can result in five Type 1 errors when a hundred studies are run. Clinical trials are only a sample of the truth, and it is unusual for Bernoulli's ninth principle to ever have application in drug development.

However, this sort of approach is used often in clinical pharmacology when looking at data from which one wants to make a regulatory claim of some sort - i.e., to convince a regulator that there is sufficient basis to grant approval to market for reasons of quality, safety, or efficacy.

In other, more experimental studies, dose is varied in different patient and volunteer populations to estimate the PK and PD properties of the drug to evaluate its potential safety and efficacy attributes. The focus here is on unbiased and precise estimation, and less on Type 1 and 2 errors and their impact on decision making.

To quantify this, we will call Θ the set of PK and PD properties we wish to estimate. Before we conduct clinical trials to characterize Θ, we will have only a rough idea (from previous experiments) or at worst, no idea, about what Θ is. Once the study or set of studies is complete, statistics will be used to quantify Θ based on the data and give the clinical pharmacologists an understanding of how the various factors involved in Θ behave.

The statistical tools of randomisation, replication, blocking, blinding, and modelling are also used in this situation, but for a different purpose. Here they are applied to ensure that the statistics give a clear idea about what Θ is (i.e., is not confounded or biased by other factors) and to meet the desired level of precision in understanding the behavior of Θ. These sorts of studies are conducted to enhance the drug company's and regulators' knowledge of the compound's properties in preparation for confirmatory trials. They do not (except in unusual circumstances) constitute sufficient evidence to permit regulators to grant market access.

1.5 Structure of the Book

Now that drug development, clinical pharmacology, and the role of statistics have been discussed, we now turn to bioequivalence. We will begin with the history of bioequivalence and an in-depth discussion of current regulatory requirements (Chapter 2). This will be followed by a lengthy chapter on the design and analysis of bioequivalence trials using 2 × 2 cross-over designs, and following this we will devote Chapter 4 to alternative designs for demonstrating bioequivalence. This is followed by Chapter 5, devoted to challenges encountered in bioequivalence studies.

There follows a brief discussion of recent proposals on alternative means of assessing bioequivalence in Chapter 6.

In subsequent chapters, we consider statistical approaches to the design and analysis of clinical pharmacology experiments to study safety (Chapter 7), QTc prolongation (Chapter 8), efficacy (Chapter 9), and population pharmacokinetics (Chapter 10).

Readers not interested with in-depth discussions of statistical theory and applications will find Chapters 1-2 and 6 most useful for their research on bioequivalence and statistics in clinical pharmacology.

CHAPTER 2

History and Regulation of Bioequivalence

Introducing Bioequivalence

It was a rainy day, and I was looking forward to another day at the Clinical Pharmacology Unit. We called it 'The Unit' for some reason. I think it was a sign of the times in the 1990s. We worked at 'The Unit'; people from FDA worked at 'The Center'; people in the CIA probably worked for 'The Agency'. You get the idea. It is good that times have changed.

It had been about a year and a half since I started working in statistics in Clinical Pharmacology, and I was starting to feel like I knew what was going on when working with the teams which were making the potential drugs and designing and performing the clinical pharmacology trials. By this time, I had worked on a couple of submissions to regulatory agencies (under supervision), had been through a regulatory audit by the FDA (they come in occasionally to check all the paperwork - as long as you follow your standard operating procedures and document what you have, this is no problem and nothing to worry about), and had figured out when lunch was.

I felt like I had it made until a ClinPharm physician and scientist came into my office that morning while I was drinking my coffee. We will call them Lenny and Denny, and they both looked like they were having a bad day. They were characters. Both of them talked a lot and at great velocity most of the time, but today they were pretty quiet. They had both been at work since 6 a.m. (clinical staff usually come in early - I think it gives them more time to make mischief) and both looked like they would rather be out in the Pennsylvania thunderstorm that was now cutting loose.

Over the monsoon, Denny filled me in on what the problem was. Lenny just nodded and groaned occasionally and looked like he wanted to go home and go back to bed. I figured he brought it on himself coming to work at 6.

In brief, one of our drugs was in the late stages of drug development. The confirmatory trials were close to finishing, and the scale-up of manufacturing to make sufficient drug to supply the marketplace had been *completed about three months ago. Everything looked pretty good - the drug was safe and well tolerated in addition to being efficacious, and we expected the Regulators to approve it once we submitted it in about six to eight months.*

The company had spent a lot of money to buy this product (we had bought it from whomever had invented it) and to develop it (estimates were in the range of what was discussed in Chapter 1) in addition to spending about five years in clinical development. It was a tremendous effort.

The problem was that the new formulation we wanted to mass produce and prepare to market clearly did not demonstrate bioequivalence to the formulation being used in the confirmatory clinical trials in a recent study. It was close, but the study did not fully meet the regulatory standard. Lenny groaned here, but I just kept drinking my coffee. I was still too new to know how bad this was. We had a quality issue in the manufacture of the drug.

This essentially meant that even if the regulators at the FDA approved the product for safety and efficacy, the company would not be able to market it. We could not (at that time) confirm that the new formula was of a sufficiently high quality to deliver the same safety and efficacy results when used in the marketplace as achieved in the confirmatory clinical trials. When Denny explained that, suddenly my coffee did not taste as good (it was always pretty bad, actually - it was free, though).

After reminding myself that I knew when lunch was supposed to be and had gotten more sleep the night before than Lenny and Denny combined (both positive factors in my view in this situation), I got a crash course in the history of bioequivalence. We then started working through the issue of how to get the quality assessment for this drug product back on track.

Biopharmaceutical statistics traditionally has focused on differentiating between products (or placebo) to provide new and enhanced treatments for the public's benefit [384]. However, this is generally expensive and time consuming (see Chapter 1) and over time steps have been taken to reduce costs and to increase supply of pharmaceutical products while maintaining the potential for innovation. One such example pertains to bioequivalence.

To call something equivalent implies a context or set of criteria for the determination of equivalence. There are several stakeholders who have say in choosing such criteria:

- *Regulatory and public-health considerations: The approach used must protect public health in that the risk of a false positive (Type 1 error) market access must be controlled at a predetermined rate.*
- *Statistical considerations: The approach should be quantifiable, accu-*

rate, precise, well understood, and should be transparent in interpretation.

- *Sponsor considerations: Using a well-designed, controlled, and reasonably sized study (or set of studies) the sponsor should be able to show the criteria have been met with a quantified chance of success.*

Various approaches to the problem of bioequivalence were considered in the 1980s through to the early 1990s, and were discarded as they failed to address one or more of the above considerations. These approaches will be considered in the remainder of this chapter.

2.1 When and How BE Studies Are Performed

Bioequivalence (BE) studies are performed to demonstrate that different formulations or regimens of drug product are similar to each other in terms of their therapeutic benefit (efficacy) and nontherapeutic side-effects (safety). They play a key and pivotal role in the drug development process by ensuring that when a patient switches to a new formulation in the marketplace, safety and efficacy will be maintained. Primarily, these studies are used in the study of solid oral dosage forms (i.e., drugs administered as a tablet or capsule when ingested), and this chapter will confine itself to discussion of this type of drug product.

When the new and old formulations use exactly the same substance (i.e., are pharmaceutically equivalent [27]) why do these studies need to be done? It is a known fact that rate and extent of bioavailability (i.e., how much drug gets into the bloodstream and is available at the site of action after one takes a dose - see Chapter 1) can be affected by very small changes in formulation. Factors like the constituent content of the formula, small changes to the lining of the formula, and by compaction into tablet (versus administration as a capsule), for example, may result in big changes in bioavailability. See [264] and [15] for examples.

Many changes are made to the formulation while Phases I to II of drug development are ongoing in clinic prior to it being approved for market access. Prior to submission to regulatory agencies and while the trials are ongoing, drug companies commonly check that these changes in formulation do not drastically change bioavailability by what are known as relative bioavailability (rel-bio) studies. These studies are primarily used by pharmaceutical sponsors of new drug entities to ensure that the formulation to be used in Phase II or in later confirmatory trials is sufficiently similar to that used in Phase I drug development and are not performed to the high requirements of true bioequivalence trials. When one wants access to the marketplace for a new formulation, a higher standard is to be met. The bioequivalence study is used to demonstrate that the formulation used in Phase III confirmatory clinical trials is sufficiently similar to the final commercial formulation to be marketed following approval.

Bioequivalence studies are primarily used by pharmaceutical sponsors of new drug entities who have conducted pivotal confirmatory trials with a specific formulation of a drug therapy but need market access for a more commercially suitable formulation (i.e., that can be mass produced). BE studies can be viewed as providing necessary and sufficient reassurance to regulators that the formulation to be marketed is the same as that used in the clinical confirmatory trials without the need to repeat the development program or to perform a therapeutic equivalence study in patients with clinical endpoints [218]. Obviously, it is impossible to repeat a drug development program with a new formulation when it is expected to last over 10 years and cost approximately a billion dollars. Such an effort is not sustainable even with modern industrial power.

Bioequivalence studies must also be performed following substantial postmarketing formulation alteration. They are also used by what is termed the 'generic' pharmaceutical industry to gain market access for formulations of established drug therapies when the patent of the original sponsor's formulation expires. When the original sponsors themselves perform a formulation change (for instance, change the site of manufacture) following approval, they often also must do a bioequivalence study to convince regulators that the new formula is safe and effective to market.

Multiple companies may produce and market similar formulations to the original marketed product following patent expiration, provided they can demonstrate bioequivalence to the original product. *Generic substitution* has thus provided a means of supplying the market with inexpensive, efficacious, and safe drug products without the need to repeat an entire clinical and clinical pharmacology development package following patent expiration.

We have now addressed when these studies are done, and we now turn to how the studies are performed. Bioequivalence studies are conducted to meet documented, legislated regulatory standards, and cross-over study designs [237], [388] are typically used to study bioequivalence. The design and application of such studies will be discussed at length in Chapter 3 but are summarized briefly here.

Bioequivalence studies are usually conducted in male and female healthy volunteer subjects. Each individual subject is administered two formulations (T=Test or R=Reference) in one of two sequences of treatments (e.g., RT and TR), see Table 2.1. R is the 'standard' and T is the 'new' formulation.

Each administration is separated by a washout period appropriate to the drug under study, see Table 2.2. This washout period consists of five

Table 2.1 *Schematic Plan of a* 2 × 2 *Cross-over Study*

Sequence Group	Period			Number of Subjects
	1	Washout	2	
1(RT)	R	—	T	$n/2$
2(TR)	T	—	R	$n/2$
R=Reference, T=Test				

half-lives between administrations. Half-life is determined by looking at the elimination (after Cmax) part of the PK concentration versus time curve (see Figure 1.3) and is simply the length of time it takes the body to eliminate one-half of the amount of whatever drug is in the body at any given time. In general, if five half-lives go by, little to no drug should be left in the systemic circulation.

Table 2.2 *Example of a Random Allocation of Sequences to Subject in a* 2 × 2 *Cross-over Design*

Subject	Sequence	Period 1	Washout period of 5 half-lives	Period 2
1	TR	T	—	R
2	RT	R	—	T
3	RT	R	—	T
.	.	.	…	.
.	.	.	…	.
.	.	.	…	.
n	TR	T	—	R

Such a design is termed a 2 × 2 cross-over [237] and is a type of design typically applied in bioequivalence trials. Of the potential list of designs (alternatives are discussed in Chapter 4) for application in bioequivalence trials, by far the most common is the 2 × 2 cross-over design (with sequences RT, TR). A potential complication for this design is that the effect of a formulation given in the first period may last into the second period, i.e., the washout period is inadequate. In the presence of such carry-over effects the interpretation of the statistics from such trials are known to be complicated [389]-[390]. When an adequate washout period is included, carry-over effects are generally considered to be negligible ([472], [473], [87], [390]). Let's go through the 2 × 2 BE design in a bit more detail.

The dose of drug substance in each formulation is pharmaceutically equivalent, and typically the formulations are not blinded (i.e., not disguised to the patient or investigator). It obviously would be difficult for a subject or clinician to bias or influence a subject's PK levels by knowing what treatment the subject received (one presumably can not change one's PK by just thinking about it).

Random allocation of subject to sequence is done here to ensure that time-related effects (i.e., period to period differences in blood sampling timings or laboratory handling of the samples, for example) can be accounted for in the analysis and are not confound with the estimate for the difference between formulations. This is an example of the practice of randomisation, and is one of the tools used to ensure bias does not creep into the study. Blood samples will be collected at predetermined, regular intervals prior to and following each dose of formulation to generate the concentration versus time curves described in Chapter 1.

Each subject serves as their own control (i.e., we can compare T to R on each subject). This is referred to as blocking and ensures that a precise measurement of the difference in formulations can be made. We will develop the model used for doing this in Chapter 3.

Replication (i.e., the number of patients assigned to each sequence) is chosen to ensure that the regulatory standards for demonstrating bioequivalence can be met. This will be a topic discussed in Chapters 3 and 4. To demonstrate equivalence in plasma concentration profiles, *rate* and *extent* of bioavailability of the drug substance in plasma must be sufficiently similar so as to meet the regulatory standard for showing that exposure of the body to the drug substance is the same between formulations [27]. For this purpose, Cmax (rate) and AUC (extent) are typically used as summary measures for the plasma concentration curves and are required to be demonstrated as equivalent under preset decision rules to achieve regulatory approval. The other pharmacokinetic endpoints discussed in Chapter 1 provide supporting information but do not directly impact approvability of the new formulation.

AUC and Cmax are looked at in this situation as surrogate markers [32] for clinical safety and efficacy. For example, if Cmax increases too much with the new formulation, this could lead to unwanted side effects. On the other hand, if it decreases too much, the drug may not be effective in treating the illness. Similar arguments apply to AUC. Hence the quality of manufacturing assessment focuses on ensuring these do not change 'too much' in the new formulation. The definition of 'too much' is quite involved and will be the subject of the next section.

Looking more closely at the endpoints we are concerned with, the pharmacokinetic endpoints AUC and Cmax are generally assumed to be, what is referred to as, log-normally distributed. A distribution is a mathematical description of the state of nature from which individual observations (like AUC and Cmax collected in our BE studies) arise. What follows is a nontechnical description of distribution theory relating to bioequivalence. Those interested in the specifics of distributional theory in this setting should review [85].

An example of a normal distribution is plotted in Figure 2.1. The density of the distribution is on the vertical axis and the corresponding AUC values are on the horizontal axis. For a given interval on the horizontal axis, the area under the curve is the probability of observing the AUC values in that interval. The larger the area of the density, the more likely are we to observe the values in the given interval. The frequency of occurrence of a lot of data in nature are well described by such a distribution. The bulk of the distribution is centered around a parameter known as the mean (μ, the measure of centrality) and is spread out to a certain extent described by the standard deviation (σ, a measure of spread). Half of the distribution falls above μ, and half falls below. Obviously, we do not know a priori what the values of μ and σ are, so we collect data and estimate them using statistics.

The role of a statistician is to use randomisation, replication, blocking, and blinding [206] in study design and proper application of models to ensure that the statistics for the parameters we are interested in are accurate and precise.

A great variety of statistical tools have been developed over the last 100 to 200 years [175]-[176] to precisely model the behavior of such normally distributed data. However, it is not uncommon for actual data not to behave themselves! AUC and Cmax data are two such examples.

Let us look at Figure 2.1 again to determine why we cannot use it directly here. Note that negative AUC or Cmax values are allowed to occur! Obviously, it doesn't make sense to use this distribution directly to describe AUC or Cmax data. One cannot physiologically have a negative blood concentration level, nor therefore a negative AUC. This situation is just not possible. The lowest they can go is 0.

There is no reason to panic, however. Statisticians (e.g., [38]) have devised a variety of ways to mathematically 'transform' non-normal data such that they can be modelled using the plethora of powerful tools involving the normal distribution which are available [258].

Westlake [452] determined that AUC and Cmax data were consistent

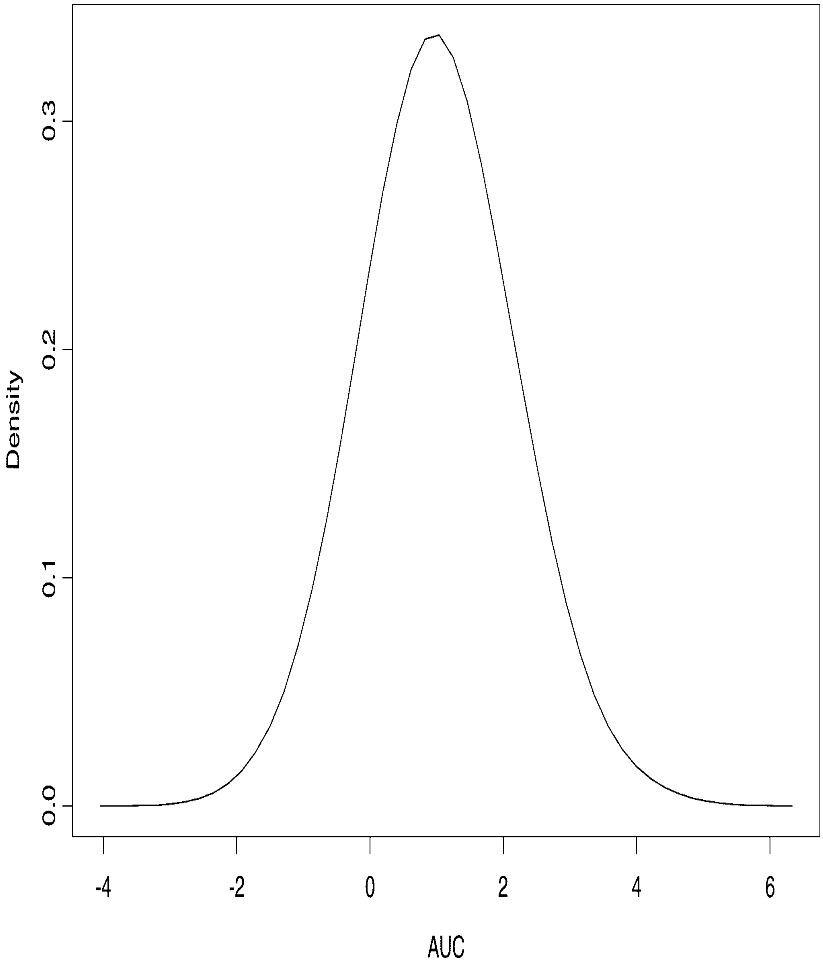

Figure 2.1 *A Normal Distribution (Mean=1, SD=1)*

with a log-normal distribution (see [302], [257], [241] for more details). This essentially means that the data are skewed such that AUC and Cmax observations must always be greater than or equal to 0. See Figure 2.2.

Mathematically, this is useful and quite convenient. If AUC and Cmax are log-normal in distribution, by taking the natural-logarithm of AUC and Cmax (i.e., by taking a mathematical *transformation*) the resulting log-transformed AUC and Cmax are normally distributed. Hence the

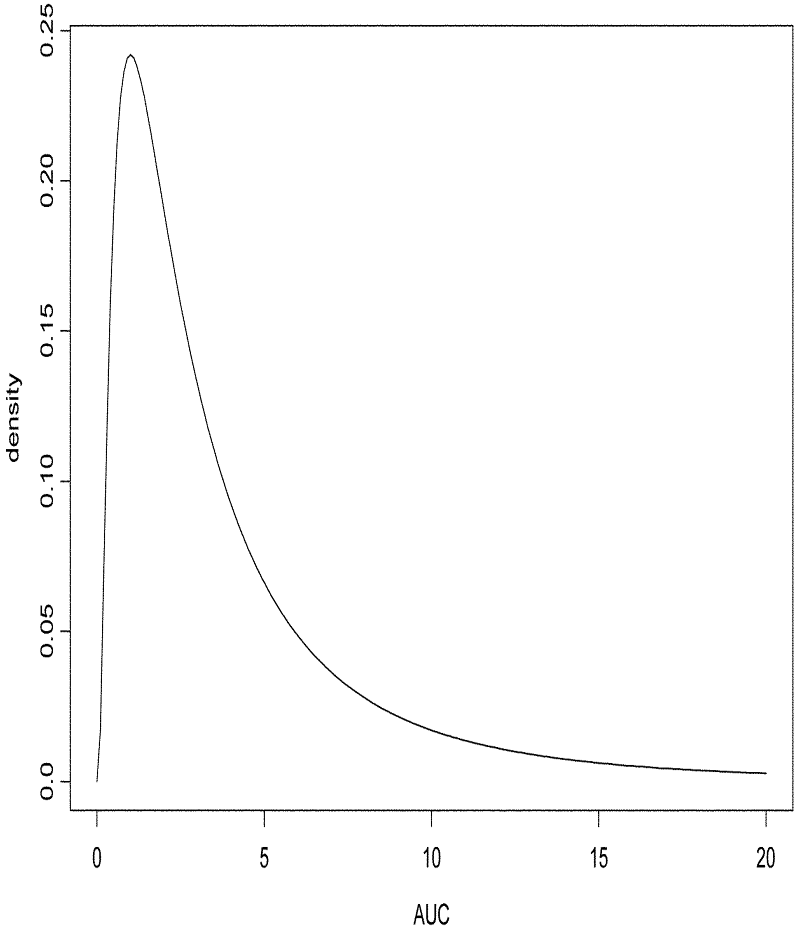

Figure 2.2 *A Log-Normal Distribution*

name - if one takes the log of a log-normal variable like AUC or Cmax, the resulting log-variable is normal in distribution.

To clarify, we take AUC as described in Figure 2.2 and recognize that the distribution is skewed and log-normally distributed. We then take the natural logarithm of the AUC values, and we get the distribution of **log**AUC plotted in Figure 2.3. Note that in Figure 2.3, the horizontal axis denotes the natural-logarithm of AUC (which is denoted mathematically as ln- or log_e-transformed AUC), which we refer to as logAUC (not AUC as in Figures 2.1 and 2.2). It is permissible for logAUC or logCmax data to have a negative value as we can always transform their value (by exponentiating) back to their original distribution where the data are always greater than or equal to 0.

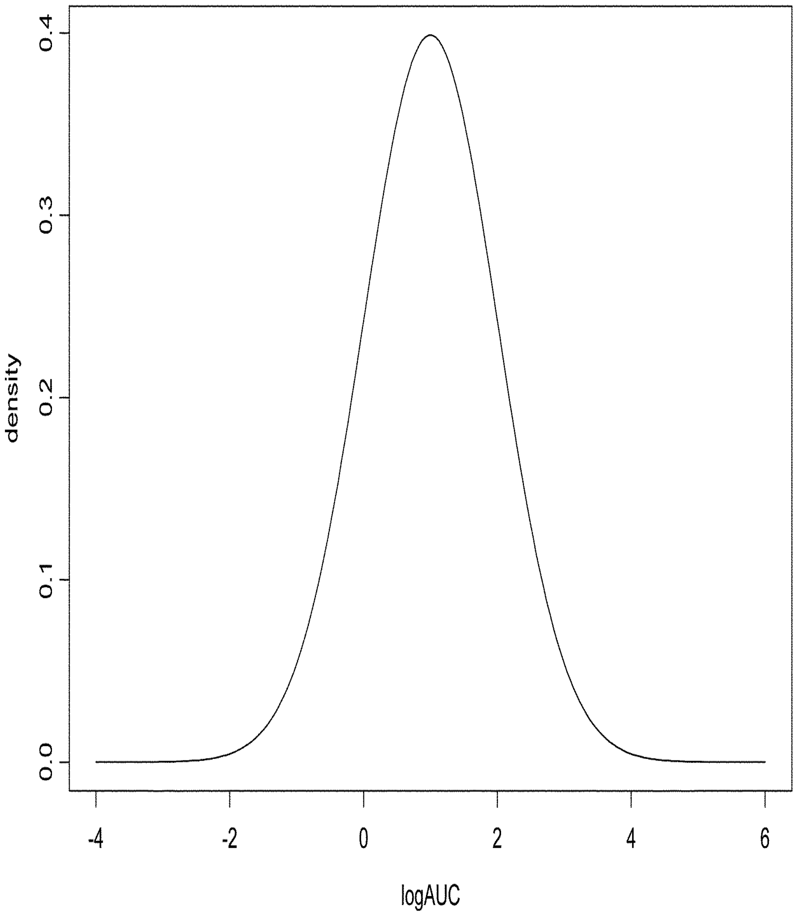

Figure 2.3 *The Normal Distribution arising from Log-Transformation for AUC Data of Figure 2.2*

To be specific for those who are interested, if AUC is log-normally distributed with mean $\exp(\mu+(1/2)\sigma^2)$ and variance $\exp(2\mu+\sigma^2)(\exp(\sigma^2)$-1), then logAUC is normally distributed with mean μ and variance σ^2 [85], [84]. This will become important later in the book as we begin modelling AUC and Cmax data.

There have been debates centered around whether AUC and Cmax are the best endpoints to use for the assessment of bioequivalence. Some findings indicate that AUC and Cmax are not always sufficient to completely demonstrate bioequivalence [359], [417], [256]; however, international regulatory authorities have depended on these endpoints since the early 1990s. Pharmacodynamic data or safety data may be required for some drug products (for an example, see [292]).

Recall that AUC is held by international regulators [61], [204], [99] to be a standard measure for extent of bioavailability. Cmax as a measure of rate of bioavailability has been found to be confounded with extent of bioavailability in studies [19] and is known to not characterize the rate of bioavailability particularly well in some situations [61].

Cmax is obviously dependent on the a priori choice of blood sampling scheme. It is known to be generally more variable than AUC and is sometimes problematic in the assessment of bioequivalence [433], [51]. Regardless of this however, Cmax has been held to be more reliable in the eyes of regulators than several alternatives [36].

Other measures of rate of absorption have been proposed in the literature such as Direct Curve Metrics [291] and Cmax/AUC [102], and indirect metrics [362]. However, simulation based assessment of alternatives has demonstrated such measures to be less desirable than the use of Cmax to date [431], [432]. Recent work in alternative measures of absorption rate such as Partial AUCs [105] is ongoing in response to workshop and regulatory considerations [393], [325] but have yet to be accepted as useful measures in bioequivalence assessment [17].

Cmax thus seems to be held as the least undesirable measure available at present for rate of bioavailability [99].

Why did something this complex ever come about? We'll go into that now.

2.2 Why Are BE Studies Performed?

In the late 1960s and 1970s, advances in chemical engineering increased the capability to create inexpensive copies of patented drug products (since termed generics). Following patent expiration, such new formulations could potentially be marketed [421].

This was desirable from a governmental perspective for public health. Such a practice would be expected to increase the supply of the products in demand in the marketplace, and thereby reduce prices for consumers. This offered substantive benefit to public health (lower costs).

However, when some pharmaceutically equivalent copies of drug products were produced, reports of therapeutic failure received a great deal of public attention in the United States. These failures included lack of desired effect (Amitriptyline, Carbamazepine, Glibenclamide, Oxytetracycline) and undesirable side effects like intoxication (Carbamazepine, Digoxin, Phenytoin). Development of a set of regulated standards for market access was necessary [360],[11]. The FDA was authorized under the 1984 Drug Price Competition and Patent Term Restoration Act to create an approval process for generic drug products.

The years following revealed increasing trends in market access for generic products [421]. For approval to market, the FDA decided to require a bioequivalence study for market access with prespecified decision rules for acceptability based on the data collected. Such studies were also required for extension of patent protection for innovators seeking to maintain market exclusivity [217].

2.3 Deciding When Formulations Are Bioequivalent

The FDA initially proposed Decision Rules (sometimes referred to as uniformity requirements) to assess bioequivalence such as the 80/20 and 75/75 rule. The 75/75 rule was defined such that 75% of subjects' individual ratios of Test to Reference, AUC or Cmax, values must be greater than or equal to the value of 0.75 for bioequivalence to be demonstrated.

While the 75/75 rule would protect against a lack of efficacy associated with decreased plasma concentrations, it obviously would not protect against undesirable side-effects potentially brought about by increased concentrations from a new formulation. Additionally, Haynes [202] established using simulation studies that the proposed 75/75 uniformity requirement was highly dependent on the magnitude of within-subject variation. Lastly, individual ratios are confounded with period effects. As these effects are known to frequently appear as significant in cross-over studies in normal healthy volunteers [376], due for example to changes in assay procedures between periods, use of the 75/75 rule criteria for bioequivalence assessment was quickly observed to be inappropriate for a large variety of drug products and was dismissed from regulatory practice.

Another idea proposed for testing bioequivalence was to simply test to see whether the formulations were different, and if the test did not demonstrate a significant difference of 20%, then one would accept bioequivalence. This was the 80/20 rule. Let μ_T (μ_R) denote the mean value of logAUC or logCmax, for T (R). Under these criteria, the study first must **not** have rejected the hypothesis H_0 that

$$H_0: \quad \mu_T = \mu_R \tag{2.1}$$

versus

$$H_1: \quad \mu_T \neq \mu_R. \tag{2.2}$$

The estimator, $\hat{\mu}_T - \hat{\mu}_R$, of $\mu_T - \mu_R$, has certain statistical properties (described in the next chapter). These may be used to derive a test-statistic and p-value to assess the above null-hypothesis H_0.

A p-value is a statistic measuring how convincing is the evidence in the data in favor of H_0. Traditionally, if its value is less than 0.05, the hypothesis H_0 is rejected in favor of its alternative H_1.

Additionally the study must have had a sufficient number of subjects to rule out the occurrence of a Type 2 error at the rate of 20% when planned to detect a clinically important difference. The use of such a procedure (known as post hoc power calculation, where power equals 1 minus the probability of a Type 2 error) is inappropriate in this context for a variety of reasons [208]. However, the clinically relevant difference was determined to be $\ln 1.25 = 0.2231$ on the log_e-scale (a 20% difference on the natural scale). See [17] for details on how this value was chosen by the FDA.

Criticisms of the 80/20 approach to bioequivalence are obvious. Absence of evidence of a significant difference does not imply evidence of absence (for more discussion see [238]). The goal of a bioequivalence s-tudy is to generate data to confirm that a difference is **not** present, not to confirm there is one. One could presumably demonstrate BE under the 80/20 rule by running a poorly conducted trial!

The statistical community had been aware, for some time, of better methods to test the hypothesis of equivalence of two treatments relative to a preset, clinically relevant goalpost. Cox [83] related Fieller's theorem [142], for the ratio of two normally distributed means, to the conditional distributions used to obtain similar regions based on traditional Neyman-Pearson theory (for the testing of hypotheses; see also [281]). Alteration of the traditional hypothesis tested in clinical trials (equations (2.1) and (2.2), above), to a framework appropriate for equivalence testing, was introduced in [95]. In this paper, Dunnett and Gent [95] compared two binomial samples relative to a prespecified goalpost Δ to assess equivalence of the responses to treatment. Westlake ([451]-[452]; for summary of work performed in the 1970s see [453]) applied similar concepts to the analysis of bioequivalence trials.

In brief, when a bioequivalence study is conducted, the confidence interval for the difference in $\mu_T - \mu_R$ is derived using a model appropriate to the data and the study design. If the confidence interval falls within prespecified goalposts, the formulations are declared bioequivalent. Implementation of the approaches proposed by Westlake [451]-[452] to the question of bioequivalence were initially assessed by Schuirmann [374] at the FDA and were subsequently adopted as the regulatory standard of choice.

This procedure was designated the 'two one-sided testing procedure' (TOST). To clarify, one hypotheses that the AUC and Cmax data in the new formulation are 'too low' (H_{01}) relative to the new formulation or also that they are 'too high' (H_{02}). If both hypotheses are rejected by the data in favor of their alternatives (H_{11} , H_{12}), then the new formulation is deemed be bioequivalent to the reference formulation.

To be specific, under this approach to inference, the usual null hypothesis was reformulated to correspond to the structure of testing the question of bioequivalence:

$$H_{01}: \quad \mu_T - \mu_R \leq -\Delta \tag{2.3}$$

versus the alternative

$$H_{11}: \quad \mu_T - \mu_R > -\Delta$$

and

$$H_{02}: \quad \mu_T - \mu_R \geq \Delta \tag{2.4}$$

versus the alternative

$$H_{12}: \quad \mu_T - \mu_R < \Delta$$

Inference was based on the use of the central t-distribution using a model in a randomized, two-period cross-over design. Summaries of the implementation of such a TOST procedure may be found in [322] and [415].

The goalpost Δ was again chosen to be equal to $\ln 1.25 = 0.2231$ (corresponding to a 20% range on the natural scale). Schuirmann subsequently refined his work in a publication in 1987 [375]. For each of the hypotheses H_{01} and H_{02} it was determined that the FDA wanted no more than a 5% chance of a Type 1 error. Recall that this means that the FDA wanted no more than a 5% chance that a study would demonstrate bioequivalence when in truth, the formulations were not bioequivalent. Examples of the application of the TOST procedure are given in Chapter 3.

Operationally, the TOST corresponds to showing that a 90% confidence interval for $\mu_T - \mu_R$, is contained in the interval $-\ln 1.25$ to $\ln 1.25$.

Blackwelder [33] and Anderson and Hauck [8] published similar work. These ideas were further developed in [186] and [363], and general approaches to the question of statistical inference were subsequently summarized under the framework of fiducial probability and inference in [319]. Practical considerations in the design and Type 2 error properties and sample size of such studies were further developed in [376].

The two one-sided testing procedure was easy to implement for nearly any cross-over study design and had the benefit of being easily interpretable in practice. As described in the last section, its regulatory and public health, statistical, and sponsor considerations were well understood.

The confidence interval provides a plausible range of values within which the true difference in formulation means can be expected to fall [187]. Note that often the results are exponentiated to the natural-scale following analysis. On the natural scale, the interval 0.80-1.25 is used to assess whether the formulations are bioequivalent for AUC and Cmax.

The ranges of plausible values as expressed by the confidence intervals were used to assess the degree of equivalence or comparability. Type 1 error was termed 'consumer' or 'regulator' risk - i.e., the risk to the regulator and consumer in making an incorrect decision, i.e., allowing market access when the application in fact should not be approved. Though often the subject of debate, the choice of $\Delta = \ln 1.25$ gave regulators an easy standard under which to assess the results of such studies.

Randomisation to sequence and definition of a washout period sufficient to negate potential residual (i.e. carry-over) effects from the previous period were established as desirable properties in bioequivalence study design. The times at which blood samples were taken was noted as being very important for proper consideration and definition of C-max, and period effects were noted as being a 'recurrent phenomenon' in cross-over designs (due to changes in sample storage, environmental conditions, or assay bias between periods - although not significant in the example provided). The use of prospectively designed, randomized cross-over designs were established as the norm for bioequivalence assessment.

Regulatory agencies have little direct interest in the Type 2 error properties of bioequivalence studies under the TOST procedure (this is typically referred to as 'sponsor's risk' in this context). The Regulator's primary concern is with the significance level at which bioequivalence can be concluded and with ensuring that the design of such studies ensures an unbiased comparison of formulations. Under Schuirmann's TOST procedure, the confidence level (α) was set at 5% per test for an overall study-wise Type 1 error rate of up to 5% [116].

The FDA recommended this in the 1992 guidance [116] and thus specified that subjects must be randomized to sequence. A general linear model (see Chapter 3) would be fitted to the log_e-transformed AUC and Cmax for demonstration of bioequivalence in a two-period cross-over design. Between- and within-subject variances were assumed to be homogeneous across formulations, and AUC and Cmax data were assumed to be log-normally distributed. In practical terms, under the 1992 FDA

Guidance, equivalence was demonstrated if the 90% confidence interval (calculated using a linear model appropriate to the study design) for $\exp(\mu_T - \mu_R)$ was contained in the interval (0.80-1.25). Different models should be applied if the study design differs from a two-period cross-over design (see Chapter 4) to construct the confidence intervals for $\mu_T - \mu_R$.

The FDA encouraged those conducting bioequivalence studies to conduct single dose studies at the maximal dose to be marketed in healthy normal subjects and to ensure an adequate washout period between study periods. AUC and Cmax were designated as the primary endpoints of interest to assess extent and rate of absorption, respectively in the 1992 FDA Guidance.

2.4 Potential Issues with TOST Bioequivalence

This *average bioequivalence* approach (so-called as it pertains to the equivalence of the means of the test and reference formulations) has safeguarded public health since its adoption [17]. However, it was not without issues.

For narrow therapeutic index drugs (for which a slight change in dose or exposure can cause a large alteration in response to treatment), bioequivalence is regarded as particularly problematic under the average bioequivalence approach [26]. Such drugs, e.g., digoxin and warfarin [78], generally exhibit low within-subject variability (i.e., within-subject coefficients of variation less than 10%). Under the average bioequivalence approach, it is possible [341] to demonstrate bioequivalence of means even in the presence of small but statistically significant changes in means - i.e., as the limits of the confidence interval for the ratio of formulation means falls within 0.80 to 1.25, bioequivalence is demonstrated; however, some confidence intervals will not contain the value 1 and thus are slightly (but significantly) different while still being bioequivalent. Such small changes in mean test to reference rate and extent of exposure are potentially clinically meaningful in a small proportion of patients [17], and some have advocated [11] special equivalence definitions for narrow therapeutic index products whereby such drugs would be held to a stricter regulatory standard (e.g., equivalence limits corresponding to a 10% range on the log_e-scale, 0.90 to 1.11).

When issues with average bioequivalence are found for a particular product (e.g., [236]), FDA typically issues a special biopharmaceutical guidance on demonstrating bioequivalence for that particular product to safeguard patients. For example, reports of therapeutic failure for the product Clozapine, an antipsychotic, were published [111]. Clozapine was granted market access following 'non-standard' bioequivalence studies mandated by FDA under biowaivers applied for by the manufacturers due to the fact that normal healthy volunteers may not be safely exposed to any dose but the lowest of Clozapine. Reports of therapeutic failure followed in the United States where uncontrolled switching in-clinic was allowed, resulting in significant costs as this condition requires hospitalization. FDA subsequently required the manufacturers of the generic formulations to perform a better bioequivalence study to maintain market access and are preparing a drug specific guidance on the topic of Clozapine bioequivalence. Examples of such drug-specific guidance include Potassium-Chloride [114], Metaproterenol and Albuterol [115], Cholestyramine [117], Phenytoin [118], Clozapine [121], and Topical Dermatologics [119].

High-variability products (with within-subject standard deviations in excess of 0.30 [34]), require sample sizes in excess of 30 subjects in order to have less than a 10% to 20% chance of a Type 2 error. Some have argued [303]-[304] that small changes in rate and extent of exposure for such products are not clinically meaningful and have advocated allowance of a less strict regulatory standard - e.g., equivalence limits corresponding to a 30% equivalence range on the log_e-scale, i.e. 0.70 to 1.43 on the natural scale. As an alternative, equivalence limits could be widened based upon the within-subject variability observed in the study [35], [372], [303]-[304] allowing such drug products easier market access.

The concept of switchability of formulations *for the individual patient* is not addressed by the average bioequivalence criterion [222]. Population means are compared, and variation between individual subjects (or patients) is factored out of the variation used to assess the distance between population means as described above. Peace [337], Anderson and Hauck [9], Hauck and Anderson [188], and Welleck [444] introduced the concept of *individual* bioequivalence. Under this approach, the question, asked is 'Can I safely and effectively switch my patient from their current formulation to another?'

Average bioequivalence is a special case of what [188] is termed *population* bioequivalence. This type of bioequivalence addresses the question, 'Can I safely and effectively start my patient on the currently approved formulation or another?' Differences in variation between formulations should also be considered when determining whether a formulation will be equally effective and safe when administering the commercial formulation of a new drug product relative to that used in clinical trials in Phase III. It is not clear in this context whether comparison of within-subject variances or total variances (so termed as the sum of between- and within-subject variance for a given formulation) is the appropriate variance for comparison between formulations, and arguments [189], [168] have been offered for both in this context.

Techniques for comparing within-subject variances in a two-period cross-over (under the assumption that between-subject variances across formulations are homogeneous) had been developed in [342] and [308]. Alternatively the total variances between formulations (between- plus within-subject variance) can be compared using a similar procedure.

Most techniques for assessment of the equality of variances assume that variance components are independent [50], [14], a condition not met in the correlated data encountered in cross-over trials. Bristol [43]-[44] developed practical maximum likelihood techniques for comparing within-subject variances in this context based on techniques discussed in [287]. Cornell [82] derived nonparametric tests of dispersion for the two-period cross-over design. Chow and Liu [74] described similar procedures, and [441] and [173]-[174] described similar procedures in publications. These techniques reduce to different transformations to assess unequal marginal scales in a bivariate normal population [247], and such comparisons were also addressed in work in [30], [97], [295], and [21]. More recent work is published in [260].

Comparisons of total- or within-subject variance between formulations can be accomplished using such procedures; however, it is known [472]-[473] that variance components are ill-characterized in cross-over studies of the size usually performed. Increasing sample size [473] can improve the precision of estimated variance components; however, it is unusual for such studies to be performed except in the case of highly variable drug products [472].

Moreover, while such procedures are theoretically and statistically viable, they are highly dependent [435] on the choice of estimation procedure. Estimates for between-subject variance can be negative under a method-of-moments based procedure or maximum-likelihood procedure [44]. Such estimates may be positively biased [109] when using restricted-maximum-likelihood based estimation procedure as would be expected in a procedure constrained in the likelihood to only permit estimates greater than or equal to zero for between-subject variances and correlation constrained to lie in the range [-1, 1] [327], [237], [88], [435].

We will not discuss the comparison of variances further in this book as such techniques are not applied in the regulatory assessment of bioequivalence. Those interested in further information on the topic should read information on this topic in [74] and the publications noted above.

Consideration of these individual and population bioequivalence ideas (and sundry others) led the FDA to form a bioequivalence working group in the mid-1990s. This body (composed of FDA representatives from clinical, scientific, and statistical disciplines) was tasked with determining whether a public health risk under the average bioequivalence approach could exist [299] and if so to determine a method or methods to evaluate bioequivalence in a manner to protect the public health. A description of the ideas under discussion may be found in [6], [189], [7], [10], [324], [166], [66], [67], and [68] but will not be discussed further here.

After considering the public comments on the preliminary draft 1997 guidance [122], the FDA reissued two draft guidances on the topic of bioequivalence in August 1999 ([125]-[126] replacing the draft guidance issued in 1997). These two guidances described when to perform a relative bioavailability, population, or individual bioequivalence study [125] for drug products in solution, suspensions, aerosols and for topical administration and for the more usual immediate-release and modified-release orally administered drug products. General guidance for study design (discussed earlier in this chapter) were provided.

The FDA acknowledged in the new draft guidance [125] that narrow therapeutic index drugs should be held to a stricter equivalence criteria than the usual 20% range required in the existing FDA guidance [116]. For these drug products, a 10% range on the log_e-scale (corresponding to an equivalence range of 0.90–1.11 on the natural scale) was required. However, this requirement was removed in the final revised FDA guidance [135].

The second draft guidance from the FDA [126] described in more detail the study design, model, and approach to statistical inference for average, population, and individual bioequivalence relative to the 1997 draft guidance, but departed from the original approach only in minor respects. Requirements for power and sample size were described in more detail in this draft guidance relative to the original 1997 draft guidance; however, the main departure was in the model used for statistical inference.

The FDA followed up in 2000 [129] with the introduction of the 'Biopharmaceutical Classification System'. Orally administered drug products are categorized based upon in vitro testing into classes I, II, III, or IV. Class I compounds, known as highly soluble and permeable in that they are quick to dissolve when ingested and are absorbed directly into the body quickly, are exempt from the requirements of demonstrating bioequivalence in a clinical study and only must demonstrate that in vitro dissolution profiles for the formulations under study are equivalent. The choice of reference product is of importance in this setting [410]. Under the BCS guidance, only Class II, III, and IV drugs are required to demonstrate in vivo bioequivalence before being granted market access.

The FDA guidance [130] finalized in October 2000 indicated that the agency would adopt the 2000 guidances [129]-[130] as final. However, following additional discussion at the 2001 Pharmaceutical Sciences Advisory Committee, the FDA provided revised final guidance [135] which removed the potential for using population and individual bioequivalence for market access. It is possible that in future the use of these criteria will be reinvestigated if the FDA determines that there is a need for such based upon observations of the marketplace. We will consider the statistical properties of alternative methods of assessing bioequivalence in a subsequent chapter.

2.5 Current International Regulation

To summarise, the debate on how to do bioequivalence trials culminated in 1987 [375] when Schuirmann's two one-sided testing method for a regulatory set goalpost of $\ln 1.25$ was introduced using the pharmacokinetic measures of AUC and Cmax as surrogate markers for efficacy and safety by the FDA.

In general, the AUC and Cmax refer to the parent compound being administered (not any metabolites produced in the body). However, under unusual circumstances, it may be important to measure metabolite AUC and Cmax also for the assessment of average bioequivalence. See [230] and [231] for more details.

The design of choice was determined to be a randomized, 2×2, two-period cross-over in normal healthy volunteers to isolate and quantify any differences in formulation, and regulatory risk was set at 5% per test. The design and analysis of cross-over studies had been extensively developed by this time [237], [388], and statistical considerations in power and sample size were described in [89].

This approach was formalized in the 1992 FDA Guidance [116] and applied to both pre- and post-marketing approvals for changes in formulation. Average bioequivalence quickly became an international standard with most nations utilizing the FDA's 1992 guidance or slight modifications to the approach.

This procedure was adopted as the standard method by European [100] and Canadian [57]-[58] regulatory authorities subsequent to finalization of the US FDA guidance in 1992 [61], [418].

Japan [232], China [70], and Australia [13] also follow this procedure (with minor changes in study design or decision rules) for the assessment of bioequivalence.

To date, the vast majority of products which have utilised this approach have not been observed to have marketplace failures in terms of their safety and efficacy profiles (see [17] for more details). Average bioequivalence testing of $\delta = \mu_T - \mu_R$ has thus been established de facto as a surrogate marker for public safety based primarily upon observation, consistency of knowledge, and replication of findings of the application of the FDA guidance [116] and less upon quantified, scientific assessment of biological plausibility and strength of association.

Average bioequivalence did, however, have the potential for issues in implementation with regard to the regulatory, statistical, and sponsor considerations discussed earlier in this chapter. One potential difficulty was regulatory in nature. The approach was concerned with testing only the formulation means and did not contain any explicit criteria pertaining to individual subjects, and it was felt that the inclusion of criteria relating to variation might address such points. Another potential area of difficulty involved both regulatory and sponsor considerations. The regulatory limits of 20% were also questioned as they might be too large for low variability products with a narrow therapeutic index, and the 20% acceptance limits created a practical difficulty for sponsors due to the large sample sizes needed to ensure a high probability of success for high variability products.

The FDA addressed this second issue presented by low variability drugs by tightening the range in some instances (e.g., for vaccines), and it was known alternative designs [435] and mixed modelling approaches [246] could be used to demonstrate average bioequivalence to address sponsor's considerations for highly variable drug products, though the statistical and regulatory considerations of such an approach were not precisely defined.

The FDA opened the discussion on the resolution of these issues with the publishing of the 1997 preliminary draft guidance [122] and significant international debate followed. This debate resolved in 2003 [135] with a final decision to continue using average bioequivalence.

Bioequivalence in practice has thus been 'harmonized' to assess the difference in means between formulations in a standard fashion throughout most of the world today.

CHAPTER 3

Testing for Average Bioequivalence

Introduction

There is nothing like a little pressure to brighten up one's day, and this one was no exception. I had arrived as usual at the clinical pharmacology unit and was at work preparing a study design proposal when my boss walked into my office.

She had run up the stairs. My office by this time (about two years after I had hired on) was one floor up and well away from my clinical colleagues who tended to be a bit noisy and nosy. The first was no problem (get some earplugs), but the second is irritating for a working statistician. They were always stopping by for just a 'peek' at the data, but they were full of questions. Answer one and a dozen more pop out. After about two years, one figures out that a little distance is not a bad thing.

After she had got her breath back, she told me that one of my colleagues from Pharmacokinetics had a bioequivalence data set that needed to be looked at 'Stat' (an expression the clinicians used all the time). I'm guessing that 'Stat' in clinician-speak means 'run the Statistics as soon as humanly possible'. I guess they like to think that we sit around twiddling our thumbs unless they shout 'Stat' repeatedly.

There is one certainty in drug development and statistics that one can depend on: the data are always late. There are always reasons that someone wanted to know the findings yesterday. Sometimes it is even a good reason!

Like the art of Statistics itself, after you get used to it, it does not bother you too much.

In any event, it was 10:30, and the results were needed by lunchtime. After making sure she meant a late lunch (she did not), I hastily pulled the code you will see later in this chapter, grabbed the data, and went to work.

We ***did*** *have a late lunch that day, by the way. Analysis of bioequivalence data is not as simple as pressing a button.*

3.1 Background

In the previous chapter we briefly introduced the 2 × 2 cross-over trial and the TOST (two one-sided testing) procedure. In this chapter we will describe in some detail how data obtained from a 2×2 trial can be used to test for Average Bioequivalence (ABE). To illustrate the analyses, we will use the data given in Table 3.1. It can be seen that data were collected on 32 subjects; 17 received the formulations in the order RT and 15 in the order TR. The original design of the trial planned for an equal number of subjects in each group. However, it is usual for such studies to overenroll to ensure that an adequate number complete the trial (without having to go to the trouble of replacing dropouts). In this case, some of the subjects did not turn up to participate in the trial. We will discuss other such practical issues of the planning of trials in Chapter 5.

Table 3.1: Example 3.1

	Sequence RT			
	AUC Period		Cmax Period	
Subject	1	2	1	2
1	2849	2230	499	436
4	2790	2864	733	416
5	2112	1744	344	48
8	1736	1882	342	437
9	1356	1175	357	240
11	1775	1585	442	286
16	2997	2237	425	332
17	1973	1778	423	407
19	1454	1297	256	348
21	2469	2023	392	480
24	1584	1855	316	373
25	4004	2449	465	625
28	1944	1593	502	326
29	1175	1147	248	221
31	1696	1801	390	350
34	1737	1655	425	319
36	2040	2199	464	384
	Sequence TR			
Subject	1	2	1	2
2	2025	2000	438	361
3	2090	1826	535	558
6	2006	1881	443	681
7	2202	1935	446	481
10	1838	1602	310	340
R=Reference, T=Test				

Table 3.1: Example 3.1

	Sequence RT			
	AUC Period		Cmax Period	
Subject	1	2	1	2
12	1898	2504	323	331
15	1129	1036	308	243
18	2014	1938	552	427
20	1900	1730	355	401
22	1763	1472	213	177
23	1678	1336	487	412
26	2271	2389	422	731
27	1986	1857	560	461
30	2519	1941	537	400
35	1560	1629	463	372
R=Reference, T=Test				

Before we proceed to test for ABE we will explore the data graphically. The main reason for using a cross-over design is to make comparisons between the two formulations 'within' each subject and as a result to eliminate any between-subject variability. Figure 3.1 is a subject-profiles plot ([237], Ch. 2) and displays the between-subject variability and the difference in response between the two formulations within each subject. The left panel displays the logAUC values and the right panel the logCmax values.

As explained in the previous chapter, the analysis is done on the log-transformed data; hence, most of the figures in this chapter will use that scale of measurement. The subject-profiles plot is constructed for each sequence by first plotting on the vertical axis, for each subject, the Period 1 and Period 2 responses against the values 1 and 2, respectively, and then joining the two responses with a line. In our plot we have replaced the axis labels for Period 1 and Period 2 with the corresponding treatment labels, so that the treatment ordering within each sequence is evident. The distance between the periods (i.e., between the R and T labels) on the horizontal axis is a matter of taste.

The expected large variation over the subjects is very evident from the vertical spread of the points in the plot. If the two formulations were identical and there was no variation in the two responses from a subject, then the plot would consist of parallel lines in a 'ladder-like' pattern. If, in addition, there was no period effect, the lines would be horizontal: if there was a period effect, all the lines would either slope upward or all the lines would slope downward. In Figure 3.1, within a sequence,

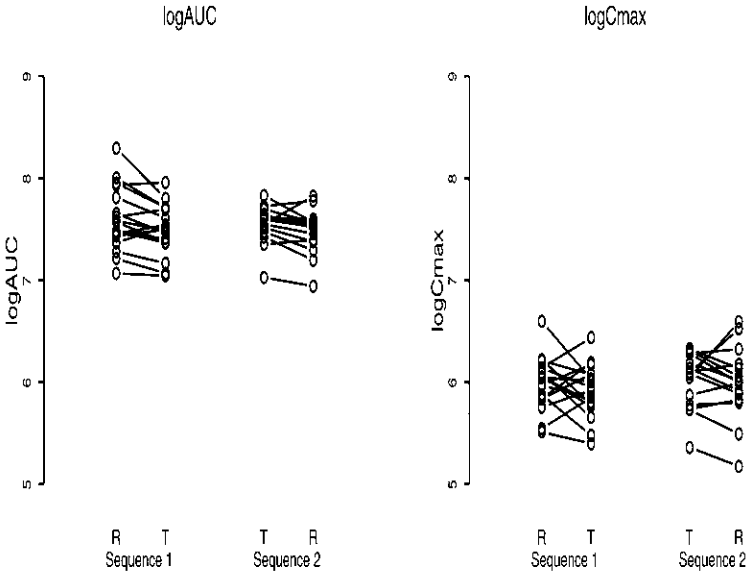

Figure 3.1 *Example 3.1: Subject Profiles Plot*

some lines go up as the formulation changes within a subject and other lines go down. It is not possible at this preliminary stage of analysis to determine whether the within-patient variability is just random noise or due to a true difference between the two formulations or between the two periods.

Another useful plot is the paired-agreement plot (see [171]). Here the Test response is plotted against the Reference response for each subject. Figure 3.2 shows, for simulated sets of data, the patterns that might be see in such a plot. These patterns correspond to (i) no difference between the two responses on a subject (Identity), (ii) a period difference in the absence of a formulation difference (Period difference), (iii) a formulation difference in the absence of a period difference (Formulation difference), and finally (iv) when there is both a period difference and a formulation difference. To emphasise the underlying pattern in each plot, we have removed the within-subject variability. For Example 3.1, the paired-agreement plots for logAUC and logCmax are shown in Figure 3.3. The patterns in the plots suggests there is a period difference but no formulation difference. In addition we can see larger within-subject variation in the logCmax values. In order to make a proper determination of any differences in response between formulations, we need to specify a statistical linear model that will allow for any systematic effects that we

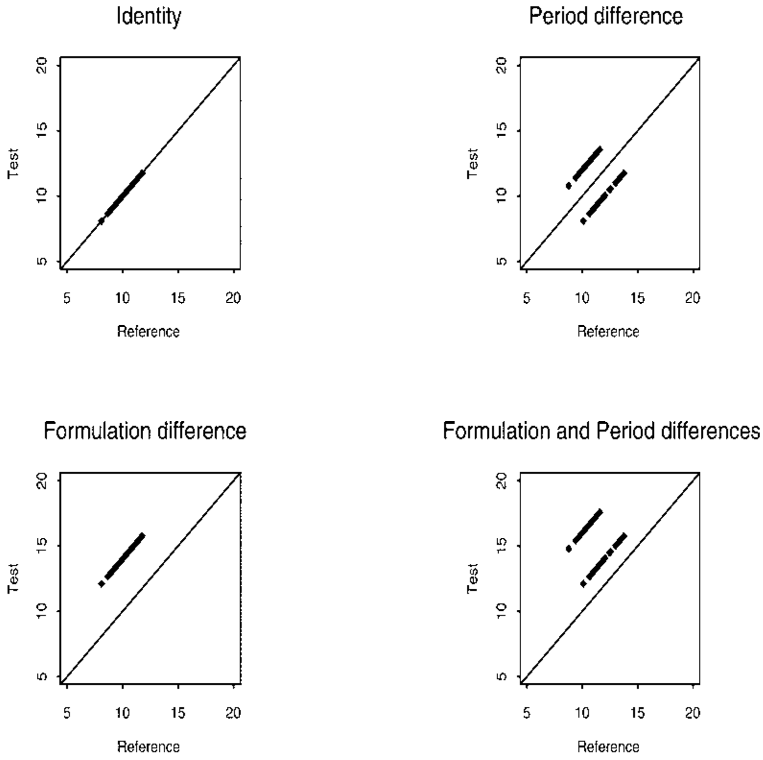

Figure 3.2 *Examples of Patterns in a Paired-agreement Plot*

believe are present in the data a priori. These systematic (fixed) effects are identified during the design phase of the trial and in our case are the sequence, formulation, and period effects. In the previous chapter reference was briefly made to so-called carry-over effects. If the effect of the formulation given in the first period is still present at the start of the second period, then we refer to that effect as the (first-order) carry-over effect of that formulation. If a long enough washout period is used to separate the two active periods (5 half-lives is recommended) then there should not be any pharmacological carry-over from the first period to the second.

Models will now be developed to generate summary statistics which account for these factors, characterize the distribution of the difference in formulation means, and to allow us to better assess the noise in the data. The essential feature of the TOST procedure is the calculation of a 90% confidence interval for $\mu_T - \mu_R$, the mean difference between

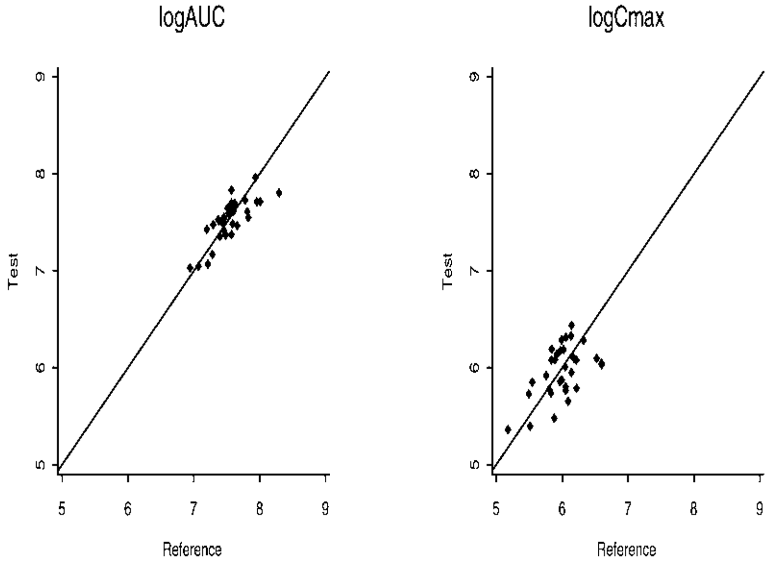

Figure 3.3 *Example 3.1: Paired-agreement Plots*

the formulations on the log scale. To calculate this confidence interval we need an estimate of $\mu_T - \mu_R$, and this can be done by specifying a statistical (linear) model for the logAUC and logCmax values observed on the subjects.

3.2 Linear Model for 2×2 Data

In order to define the linear model let y_{ijk} denote the response (i.e., logAUC or logCmax) in period j on subject k in sequence group i, where $i = 1, 2$, $j = 1, 2$, $k = 1, 2, \ldots, n_i$, and n_i is the number of subjects in group i. The total number of subjects in the trial is $n = n_1 + n_2$. The systematic effects we anticipate are due to the periods and formulations. As the subjects are allocated randomly to the two groups, there should be no sequence effect (i.e., a significant difference in mean response between the two sequence groups). However, it is traditional to include such an effect and we will do so here. The notation we will use is that μ denotes the overall mean response, τ_R and τ_T are the formulation effects, π_1 and π_2 are the period effects and γ_1 and γ_2 are the sequence effects. The fixed effects model (i.e., the systematic effects) for each of the four group-by-period response combinations is displayed in Table 3.2. As we will explain later, the difference in carry-over effects, if any, between

Table 3.2 *Fixed Effects in the Linear Model for the* 2×2 *Design*

Group	Period 1	Period 2
1(RT)	$\mu + \tau_R + \pi_1 + \gamma_1$	$\mu + \tau_T + \pi_2 + \gamma_1$
2(TR)	$\mu + \tau_T + \pi_1 + \gamma_2$	$\mu + \tau_R + \pi_2 + \gamma_2$

Table 3.3 *Fixed Effects: Alternative Parametrization for the* 2×2 *Design*

Group	Period 1	Period 2
1(RT)	$\mu_R + \pi_1 + \gamma_1$	$\mu_T + \pi_2 + \gamma_1$
2(TR)	$\mu_T + \pi_1 + \gamma_2$	$\mu_R + \pi_2 + \gamma_2$

the formulations is aliased with (i.e., completely mixed up with) any difference between the sequence effects, so including sequence effects in our model does have some potential benefits which we will explore later. As regards the parameters themselves, they are all defined with reference to an overall mean response parameter μ. The result of moving from R to T, for example, is to cause an increase or a decrease in response relative to the overall mean. Consequently, as it is only the size of the increase or decrease that needs to be accounted for, two different formulation parameters, τ_R and τ_T, are not needed. To remove this redundancy a constraint is typically applied such as $\tau_R + \tau_T = 0$. The result of this is that we can refer to $\mu_R = \mu + \tau_R$ and $\mu_T = \mu + \tau_T$ as the means for formulation R and T, respectively. This alternative parametrization is displayed in Table 3.3. For exactly the same reasons a constraint is also placed on the period and sequence parameters, e.g., $\pi_1 + \pi_2 = 0$ and $\gamma_1 + \gamma_2 = 0$. The choice of constraint is not unique and we could have chosen $\tau_R = 0$ and $\pi_1 = 0$, for example. What is important to remember is that, although the choice of constraint is arbitrary, the difference $\mu_T - \mu_R$ is uniquely identified.

Coming now to the 'random-effects' part of our model, we need to allow for the variation between patients that was so evident in Figures 3.1 and 3.3 and for any 'residual' random variation that is unexplained by the rest of the terms in the model. This is done by introducing two random variables: $\xi_{k(i)}$, to allow for variation between subjects and ε_{ijk} to

Table 3.4 *Random Effects in the Linear Model for the* 2×2 *Design*

Group	Period 1	Period 2
1(RT)	$\xi_{k(1)} + \varepsilon_{11k}$	$\xi_{k(1)} + \varepsilon_{12k}$
2(TR)	$\xi_{k'(2)} + \varepsilon_{21k'}$	$\xi_{k'(2)} + \varepsilon_{22k'}$

allow for unexplained variation between the two responses on the same subject. The random effects are displayed in Table 3.4 for a typical subject k in Group 1(RT) and for a typical subject k' in Group 2(TR). Dropping the distinction between k and k', we assume that $\xi_{k(i)}$ and ε_{ijk} are independent random variables such that $\mathrm{E}(\xi_{k(i)}) = 0$, $\mathrm{Var}(\xi_{k(i)}) = \sigma_B^2$, $\mathrm{E}(\varepsilon_{ijk}) = 0$ and $\mathrm{Var}(\varepsilon_{ijk}) = \sigma_W^2$, where σ_B^2 is the between-subject variance and σ_W^2 is the within-subject variance. E denotes the expected-value (i.e., population mean) for a given parameter, and Var denotes its variance. We also assume that the $\xi_{k(i)}$ are independent among themselves and that the ε_{ijk} are independent among themselves. The complete model for y_{ijk} is then:

$$y_{ijk} = \mu_{d[i,j]} + \pi_j + \gamma_i + \xi_{k(i)} + \varepsilon_{ijk}, \tag{3.1}$$

where $d[i,j] = R$ or T and identifies the formulation in period j of sequence i.

We note that the variance of a response on subject k in group i in period j is:

$$\sigma^2 = \mathrm{Var}(y_{ijk}) = \mathrm{Var}(\xi_{k(i)} + \varepsilon_{ijk}) = \sigma_B^2 + \sigma_W^2. \tag{3.2}$$

The covariance between two responses on the same subject is

$$\mathrm{Cov}(y_{i1k}, y_{i2k}) = \mathrm{Cov}(\xi_{k(i)} + \varepsilon_{i1k}, \xi_{k(i)} + \varepsilon_{i2k}) =$$
$$\mathrm{Cov}(\xi_{k(i)}, \xi_{k(i)}) = \mathrm{Var}(\xi_{k(i)}) = \sigma_B^2.$$

Hence, the correlation between two responses on the same subject is:

$$\rho = \mathrm{Corr}(y_{i1k}, y_{i2k}) = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}. \tag{3.3}$$

Returning now to the estimation of $\mu_T - \mu_R$, let $\bar{y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}$ denote the mean response of the subjects in period j in sequence group i.

$$\text{For Group 1: } \mathrm{E}(\bar{y}_{11.} - \bar{y}_{12.}) = \pi_1 - \pi_2 + \mu_R - \mu_T.$$
$$\text{For Group 2: } \mathrm{E}(\bar{y}_{21.} - \bar{y}_{22.}) = \pi_1 - \pi_2 + \mu_T - \mu_R.$$

Table 3.5 *Example 3.1: Groups-by-Periods Means (sample size in brackets)*

		logAUC	
Group	Period 1	Period 2	Mean
1(RT)	$\bar{y}_{11.} = 7.60(17)$	$\bar{y}_{12.} = 7.50(17)$	$\bar{y}_{1..} = 7.55$
2(TR)	$\bar{y}_{21.} = 7.55(15)$	$\bar{y}_{22.} = 7.48(15)$	$\bar{y}_{2..} = 7.51$
Mean	$\bar{y}_{.1.} = 7.58$	$\bar{y}_{.2.} = 7.49$	$\bar{y}_{...} = 7.53$
		logCmax	
1(RT)	$\bar{y}_{11.} = 5.99(17)$	$\bar{y}_{12.} = 5.91(17)$	$\bar{y}_{1..} = 5.95$
2(TR)	$\bar{y}_{21.} = 6.02(15)$	$\bar{y}_{22.} = 5.99(15)$	$\bar{y}_{2..} = 6.01$
Mean	$\bar{y}_{.1.} = 6.01$	$\bar{y}_{.2.} = 5.95$	$\bar{y}_{...} = 5.98$

Hence,

$$\mathrm{E}\left\{\frac{1}{2}\left[(\bar{y}_{21.} - \bar{y}_{22.}) - (\bar{y}_{11.} - \bar{y}_{12.})\right]\right\} = \mu_T - \mu_R.$$

That is,

$$\hat{\mu}_T - \hat{\mu}_R = \frac{1}{2}(\bar{y}_{21.} - \bar{y}_{22.} - \bar{y}_{11.} + \bar{y}_{12.}) \tag{3.4}$$

and

$$\mathrm{Var}(\hat{\mu}_T - \hat{\mu}_R) = \frac{1}{4}\left[\frac{\sigma_W^2}{n_1} + \frac{\sigma_W^2}{n_1} + \frac{\sigma_W^2}{n_2} + \frac{\sigma_W^2}{n_2}\right] = \frac{\sigma_W^2}{2}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]. \tag{3.5}$$

If $n_1 = n_2 = n/2$, then

$$\mathrm{Var}(\hat{\mu}_T - \hat{\mu}_R) = \frac{\sigma_W^2}{2}\left[\frac{2}{n} + \frac{2}{n}\right] = \frac{2\sigma_W^2}{n}. \tag{3.6}$$

If $\hat{\sigma}_W^2$ is an estimate of σ_W^2 on $n-2$ degrees of freedom (d.f.) and $t_{0.95}(n-2)$ is the upper 95% percentile of the $t-$distribution on $n-2$ d.f., the 90% confidence interval for $\mu_T - \mu_R$ is

$$\hat{\mu}_T - \hat{\mu}_R \pm t_{0.95}(n-2)\sqrt{\frac{\hat{\sigma}_W^2}{2}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}. \tag{3.7}$$

The groups-by-periods means for Example 3.1 are given in Table 3.5.

Finally, we display a plot that is directly linked to the linear model and displays information on both the formulation difference within subjects

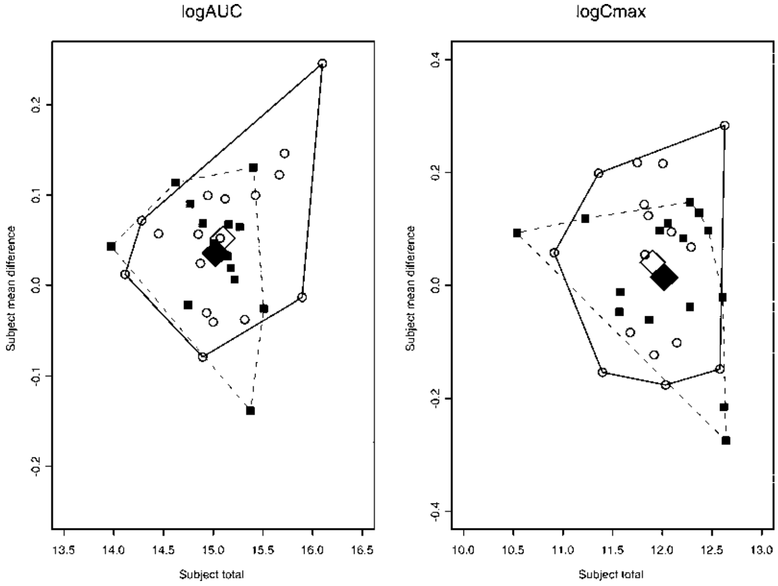

Figure 3.4 *Example 3.1: Mean Differences versus Totals Plot*

and their variability. In this 'Mean Differences versus Totals' plot we plot for each subject k in Group i, the mean difference $d_{ik} = (y_{i1k} - y_{i2k})/2$ against the total $t_{ik} = y_{i1k} + y_{i2k}$ (see [237] and [171]). The resulting plot is given in Figure 3.4, where open symbols are used for the subjects in Group 1. The two large diamonds on each plot indicate the position of the centroids $[(\bar{t}_{1.}, \bar{d}_{1.}), (\bar{t}_{2.}, \bar{d}_{2.})]$. The vertical difference between the centroids within a plot is the value of $\hat{\mu}_T - \hat{\mu}_R$. We can see that for both logAUC and logCmax the centroids are close together, suggesting that T and R might be ABE. The solid and dashed lines in each plot give the positions of the convex hulls, one for each group. The convex hull connects the 'outermost' points in a group, and is a useful way of displaying the variation in the d_{ik} and t_{ik}. There is an impression that variability is higher in Group 1 for both logAUC and logCmax. The usefulness of plotting the subject totals is that the difference $\bar{t}_{1.} - \bar{t}_{2.}$ is an estimate of the difference in the carry-effects of T and R (see [237], Chapter 2). For BE trials with an adequate washout period this difference should be zero. Testing for a difference in carry-over effects in the RT/TR design is problematic and we do not recommend it. We say more on carry-over effects in Section 3.4.

3.3 Applying the TOST Procedure

The SAS code to fit Model (3.1) and calculate the 90% confidence interval is given in the following boxes. An edited version of the output produced is given in the boxes immediately following the SAS code. The results of applying the TOST procedure are given in Table 3.6.

The ABE limits are (-0.2231, 0.2231) on the log scale and (0.8,1.25) on the original scale. Clearly the confidence intervals for both AUC and Cmax are well with the ABE limits and so T and R an be declared equivalent based on the ABE criterion.

A graphical representation of the results is given in Figure 3.5, where the density of the normal distribution based on the fitted mean and standard error for each of logAUC and logCmax are plotted along with the ABE limits. Both densities are well within the limits indicating that T and R are average bioequivalent. It is also apparent that the density for logCmax is wider than that for logAUC indicating that Cmax is a more variable metric than AUC in this particular trial.

We note that the estimated correlation between the two responses on the same subject can be estimated from the SAS output. For logAUC $\hat{\sigma}_B^2 = 0.052$ and appears under the output heading of ‘Covariance Parameter Estimates’ in the row labelled as ‘SUBJECT(SEQUENCE)’. For logAUC, $\hat{\sigma}_W^2 = 0.011$ and appears in the row labelled ‘Residual’. The estimated correlation coefficient for logAUC is then $\hat{\rho}_{\text{logAUC}} = 0.052/(0.052 + 0.011) = 0.83$. For logCmax the corresponding value is $\hat{\rho}_{\text{logCmax}} = 0.045/(0.045 + 0.038) = 0.54$. There is a higher level of total variability for logCmax as compared to logAUC, as was already concluded from Figures 3.3 and 3.4.

ABE Example 3.1 - SAS `proc mixed` *Code:*

```
data ABEexample1;
input subject sequence$
formulation$ period AUC CMAX;
logauc=log(AUC);
logcmax=log(CMAX);
datalines;
 1 RT R 1 2849 499
 1 RT T 2 2230 436
 .  . . .   .   .
 .  . . .   .   .
 .  . . .   .   .
35 TR R 2 1629 372
35 TR T 1 1560 463
;
run;
```

ABE Example 3.1 - SAS `proc mixed` *Code, continued:*

```
proc mixed data=ABEexample1;
class sequence subject period
formulation;
model logauc=sequence period
formulation/ddfm=kenwardroger;
random subject(sequence);
lsmeans formulation/pdiff cl alpha=0.1;
estimate 'ABE for logAUC' formulation -1 1;
run;

proc mixed data=ABEexample1;
class sequence subject period formulation;
model logcmax=sequence period formulation/
ddfm=kenwardroger;
random subject(sequence);
lsmeans formulation/pdiff cl alpha=0.1;
estimate 'ABE for logCmax' formulation -1 1;
run;
```

ABE Example 3.1 - Edited SAS Output:

```
Log AUC
Covariance Parameter Estimates
Cov Parm              Estimate
SUBJECT(SEQUENCE)      0.0516
Residual               0.0110
                Standard
Effect  Estimate Error   DF
T-R    -0.0166   0.0263  30
Alpha  Lower   Upper
0.1   -0.0612  0.0280
```

ABE Example 3.1 - Edited SAS Output:

```
Log Cmax
Covariance Parameter Estimates
Estimate SUBJECT(SEQUENCE) 0.04528
Residual                   0.03835
                Standard
Effect  Estimate Error   DF
T-R    -0.0269  0.0490  30
Alpha  Lower   Upper
0.1   -0.1102  0.0563
```

As a final summary we display the confidence intervals on the natural

Table 3.6 *Example 3.1: TOST Procedure Results*

Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% Confidence Interval
AUC	-0.0166	(-0.0612, 0.0280)
Cmax	-0.0269	(-0.1102, 0.0563)
Endpoint	$\exp(\hat{\mu}_T - \hat{\mu}_R)$	90% Confidence Interval
AUC	0.98	(0.94, 1.03)
Cmax	0.97	(0.90, 1.06)

scale alongside a plot of the ratios T:R for each of AUC and Cmax in Figure 3.6. We note that, for Cmax especially, there are many subjects that have ratios outside the ABE limits of (0.8, 1.25). This example highlights that fact that to be equivalent on the ABE criterion it is only necessary to show that the means of T and R do not differ to a significant extent.

Before leaving this section, we demonstrate that the confidence interval testing approach we have used is equivalent to the alternative version of the TOST procedure that requires the testing of two, one-sided hypotheses:

$$H_{01}: \quad \mu_T - \mu_R \leq -\Delta \tag{3.8}$$

versus the alternative

$$H_{11}: \quad \mu_T - \mu_R > -\Delta$$

and

$$H_{02}: \quad \mu_T - \mu_R \geq \ \Delta \tag{3.9}$$

versus the alternative

$$H_{12}: \quad \mu_T - \mu_R < \ \Delta.$$

Here, it will be recalled, $\Delta = \ln(1.25) = 0.2231$.

We first consider logAUC. The value of the t-statistic for testing (3.8) is $t_{01} = (-0.0166 + \Delta)/0.0263 = 7.85$ on 30 d.f. and the value for testing (3.9) is $t_{02} = (-0.0166 - \Delta)/0.0263 = -9.11$ on 30 d.f. Clearly both of these null hypotheses would be rejected at the 5% level on a one-sided test. For logCmax, the story is similar with a value of 4.00 for testing (3.8) and a value of -5.10 for testing (3.9), both on 30 df.

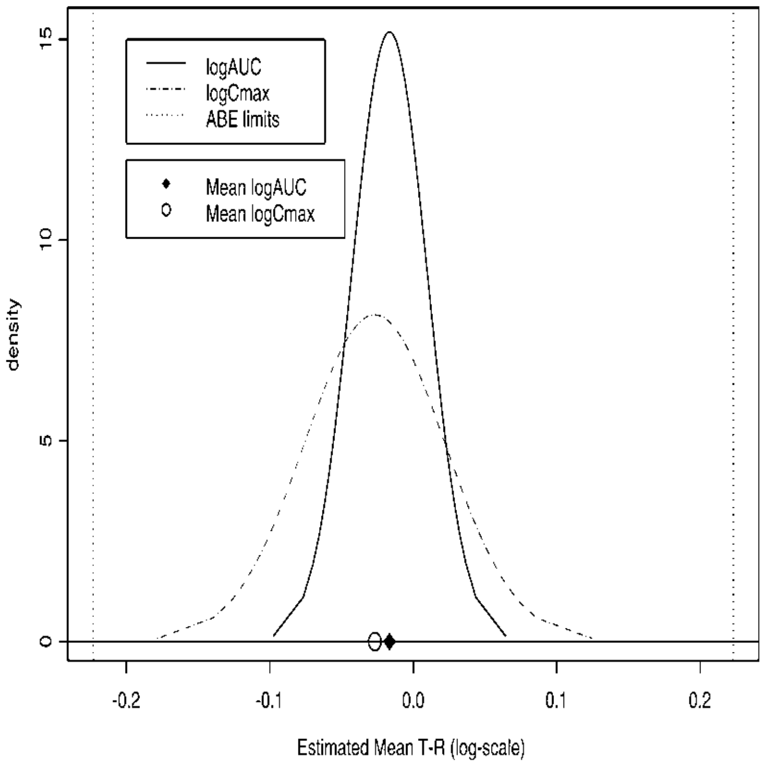

Figure 3.5 *Example 3.1: Fitted Normal Densities for* $\hat{\mu}_T - \hat{\mu}_R$

3.4 Carry-over, Sequence, and Interaction Effects

We now return to consider the other potential effects that might be present in our data, namely carry-over and formulation-by-period interaction. The nature of the carry-over effects was described in the Introduction, so we do not repeat that here. The interaction effect, however, is something we have not yet considered. Our current Model (3.1) assumes that the difference between μ_T and μ_R is the same in Period 2 as it is in Period 1. This is the situation when there is no formulation-by-period interaction. The presence of such an interaction implies that the size of the formulation difference in Period 1 is not the same as its size in Period 2. Figure 3.7 contains four examples of a groups-by-periods plot ([237], Ch. 2) which displays the four group-by-period means $\bar{y}_{11.}, \bar{y}_{12.}, \bar{y}_{21.}$, and $\bar{y}_{22.}$. They are given in two versions and each for the cases of no-interaction and interaction. Let us look first at the Version 1 plots. If we assume for the moment that there is no random variation,

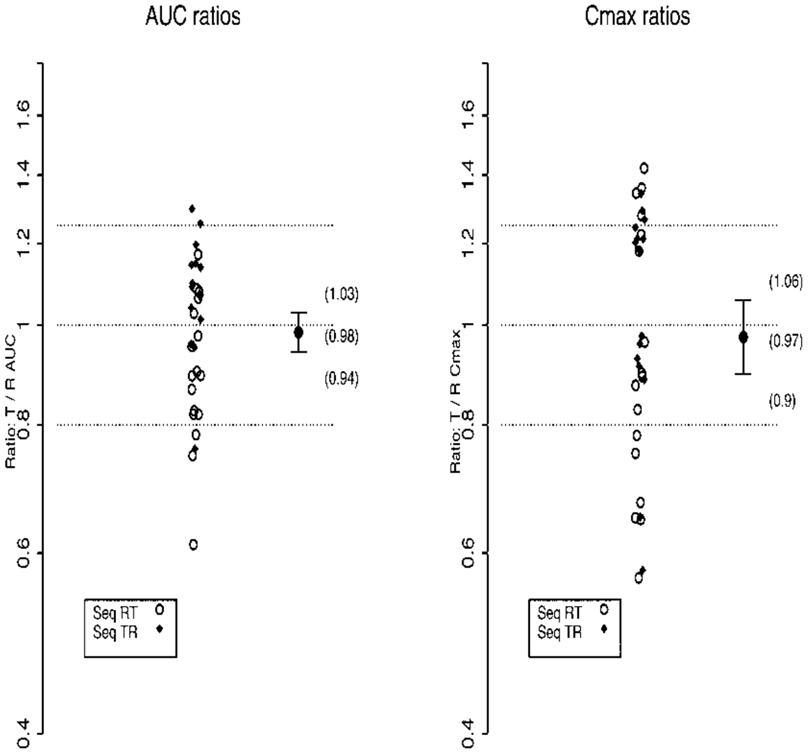

Figure 3.6 *Example 3.1: 90% Confidence Intervals for* $\exp(\hat{\mu}_T - \hat{\mu}_R)$.

i.e., the plotted points refer to the true mean values, then the upper left-hand plot is a case where there is no interaction and as a consequence the lines cross at a point midway between Period 1 and Period 2. The lower left-hand plot is a case where there is interaction, and the lines cross at a position that is not midway between the two period labels. Deciding quickly if the crossing point is midway or not may not be easy and so Version 2 offers an alternative. Here the points are in the same positions, but an alternative way of connecting them has been used. If the lines are parallel then there is no interaction. The reader can decide which version, if any, they find useful. In all the upper plots a period effect is evident that gives a lower response in Period 2. In the presence of random variation, we will not see parallel lines even in the absence of any interaction: a statistical test of significance will be required to determine if there is any evidence of an interaction. If this is the case, then the fixed effects part of our model, as displayed in Table 3.2, needs

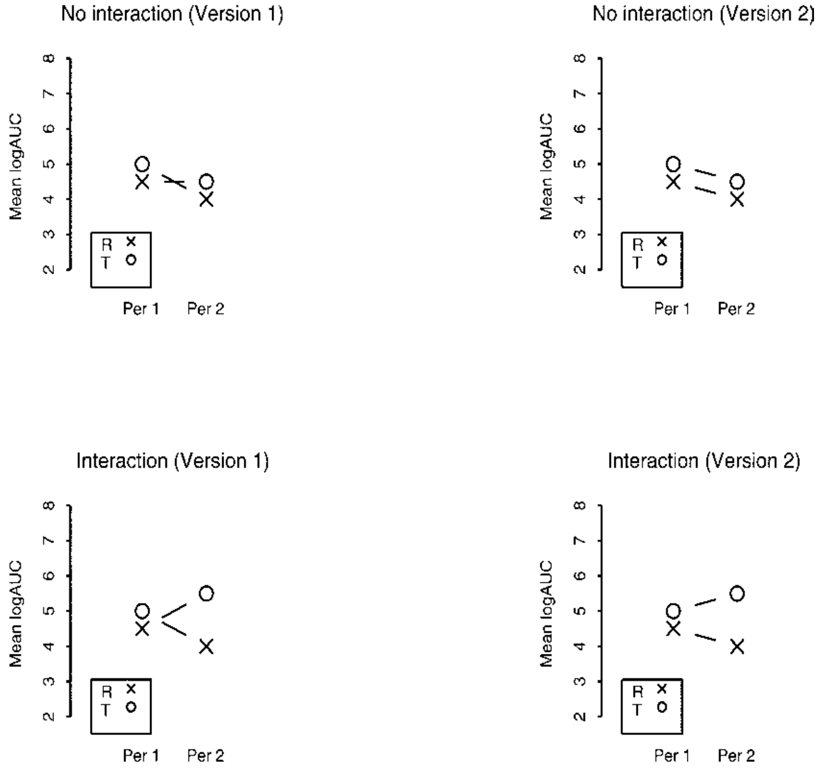

Figure 3.7 *Groups-by-Periods Plot*

to be enlarged to that given in Table 3.7, where there are four new (interaction) parameters $(\tau\pi)_{d[i,j],j}, i = 1, 2, j = 1, 2$ and where $d[i, j] = R$ or T. (Note, we have omitted the sequence parameters for reasons which will be explained shortly.) The inclusion of these parameters implies that a response observed in Group i, Period j under formulation R or T is not just the sum of the individual effects of formulation $d[i, j]$ and period j. However, as already mentioned in Section 3.2, our linear model is over-parameterized and we need to apply constraints to remove the redundancy. The constraints on the formulation and period parameters have already been described. For the interaction parameters we assume: $(\tau\pi)_{R1} = (\tau\pi) = -(\tau\pi)_{R2} = -(\tau\pi)_{T1} = (\tau\pi)_{T2}$. The model containing the interaction parameters is displayed in Table 3.8. From this table it is also clear that the sequence parameter for Sequence $1(-\gamma)$ and the interaction parameter $(\tau\pi)$ are interchangeable in Group 1 and (γ) and $-(\tau\pi)$ are interchangeable in Group 2. Hence, the statistical test for

Table 3.7 *Fixed Effects Model Including Interactions for a* 2×2 *Design*

Group	Period 1	Period 2
1(RT)	$\mu_R + \pi_1 + (\tau\pi)_{R1}$	$\mu_T + \pi_2 + (\tau\pi)_{T2}$
2(TR)	$\mu_T + \pi_1 + (\tau\pi)_{T1}$	$\mu_R + \pi_2 + (\tau\pi)_{R2}$

Table 3.8 *Fixed Effects Model Including Interactions for a* 2×2 *Design: After Applying Constraints*

Group	Period 1	Period 2
1(RT)	$\mu_R - \pi + (\tau\pi)$	$\mu_T + \pi + (\tau\pi)$
2(TR)	$\mu_T - \pi - (\tau\pi)$	$\mu_R + \pi - (\tau\pi)$

a group difference is identical to the test for a non-zero formulation-by-period interaction. In this situation we say that the sequence and interaction effects are aliased. The same can be said about the carry-over difference. To see this we need to apply a different constraint to the four interaction parameters (recall the choice of constraint is arbitrary). Table 3.9 displays the model with carry-over effects. There is no carry-over effect in Period 1 and $\lambda_R(\lambda_T)$ denotes the carry-over effect of formulation $R(T)$. If we apply the constraint $\lambda_R = -\lambda = -\lambda_T$, then the model is as displayed in Table 3.10. If we return to Table 3.7 and apply the constraints: $(\tau\pi)_{R1} = (\tau\pi)_{T1} = 0$, $(\tau\pi)_{R2} = -\lambda$ and $(\tau\pi)_{T2} = \lambda$, we will reproduce Table 3.10. In other words the carry-over effects and the interaction effects are aliased.

We may now be tempted to test for a nonzero formulation-by-period interaction (or carry-over difference or group difference). However, such a test is pointless and has undesirable side effects. The reasons why this is the case were first given by [150] and subsequently described and discussed thoroughly by Senn (see [383], [384], [389], [390]), for example) and [237]. We therefore do not consider or recommend such testing for trials that use the RT/TR design.

For completeness we show, in Figure 3.8, the groups-by-periods plots for Example 3.1, where the style of Version 1 has been used. Although

Table 3.9 *Fixed Effects Model Including Carry-over Effects for a* 2×2 *Design*

Group	Period 1	Period 2
1(RT)	$\mu_R + \pi_1$	$\mu_T + \pi_2 + \lambda_R$
2(TR)	$\mu_T + \pi_1$	$\mu_R + \pi_2 + \lambda_T$

Table 3.10 *Fixed Effects Model Including Carry-over for a* 2×2 *Design: After Applying Constraints*

Group	Period 1	Period 2
1(RT)	$\mu - \tau - \pi$	$\mu + \tau + \pi - \lambda$
2(TR)	$\mu + \tau - \pi$	$\mu - \tau + \pi + \lambda$

there is a suggestion of an interaction, it is unlikely that such a small effect, if in fact it is present, could be detected against a background of large between-subject variability. In addition, we have already cautioned against testing for such an interaction. In terms of ABE, although the logCmax means in Period 2 show more of a difference between R and T than the other comparisons, this difference is itself not large.

3.5 Checking Assumptions Made about the Linear Model

No statistical analysis of data is complete without some checks on the assumptions that were made when the model was specified. Our model, it will be recalled is as defined in (3.1). The main assumptions were that after allowing for the systematic (i.e., fixed) effects, the between-subject variability and the within-subject variability can be modelled by normal distributions. A simple graphical test of whether a set of values is a sample from a normal distribution is the normal probability (or Q-Q) plot. The values of most interest to us are the within-subject residuals, i.e., the estimates of the ε_{ijk}. We will denote the residual for the kth subject in sequence i and period j as r_{ijk}. It is defined as $r_{ijk} = y_{ijk} - \hat{y}_{ijk}$, where $\hat{y}_{ijk}$ is the value our model predicts (using our given data) for the response of the kth subject in sequence i and period j.

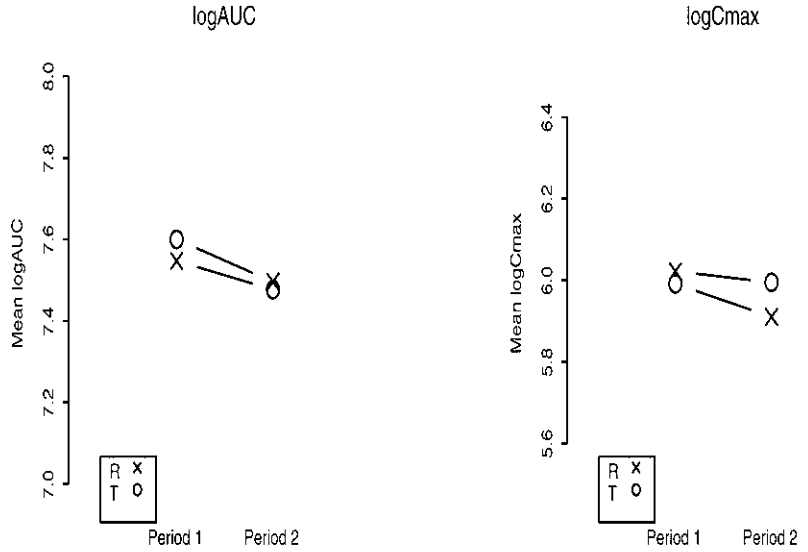

Figure 3.8 *Example 3.1: Groups-by-Periods Plot*

Because our model measures everything relative to the grand mean (μ), the two residuals on the same subject add to zero, i.e., ($r_{i1k} + r_{i2k} = 0$). Hence, when testing the residuals for normality, we need only use the residuals from one of the periods.

Figure 3.9 displays the Q-Q plots for the studentized residuals corresponding to logAUC and logCmax. Identified on the plots are the two most extreme residuals in each plot. The studentized residuals are the raw residuals (r_{ijk}) divided by their estimated standard error. The standardization is necessary because Var(r_{ijk}) is not a constant. If the plotted data are truly normally distributed, the plotted points should lie on or close to a straight line. We can see that this is mostly true in Figure 3.9, except for the logAUC values of two subjects (12 and 25). A more formal test of normality is one due to Shapiro and Wilk [397]. For logAUC the p-value for this test is 0.497 and for logCmax is 0.314. There is no evidence to suggest the studentized residuals are not normally distributed. The responses with the largest studentized residuals (in absolute value) may be outliers. These are values that are typically greater than 3 in value. There is no evidence that our extreme residuals are outliers.

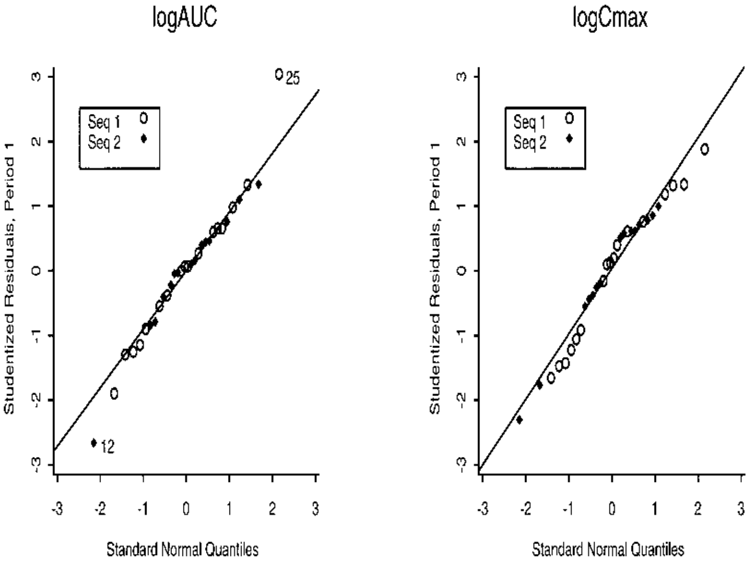

Figure 3.9 *Example 3.1: Normal Probability Plots*

3.6 Power and Sample Size for ABE in the 2×2 Design

In order for an ABE trial to meet its objectives, it should have a good chance of deciding that T and R are average bioequivalent when that is, in fact, the true state of nature. Expressed in statistical terminology, the trial must have sufficient power to reject the two null hypotheses of non-equivalence, when T and R are average bioequivalent. Power is the probability of rejecting the two null hypotheses, when they are false, and is usually chosen to be 0.90. Mathematically, power equals 1 minus the probability of a Type 2 error.

No adjustment is made for multiplicity [191] and the larger variance of logAUC or logCmax is used in the power sample size calculations. It is generally the case that logCmax is more variable than logAUC, as illustrated in the previous example. Such practical issues in determining the sample size of ABE trials are considered in more detail in Chapter 5.

As already explained, when using the TOST procedure to determine ABE we test each of the following two null hypotheses at a significance level of α, where $\Delta = \ln 1.25$. If both are rejected we conclude that for the metric being used (logAUC or logCmax) that T and R are ABE.

$$\begin{aligned} H_{01} :&\mu_T - \mu_R \leq -\Delta \\ H_{02} :&\mu_T - \mu_R \geq \quad \Delta. \end{aligned}$$

In practice it is convenient to do this using a $100(1-2\alpha)\%$ two-sided confidence interval. However, in order to calculate the power of the TOST procedure, we will stay within the hypothesis testing framework. The power will be calculated for a 2 × 2 cross-over trial with n subjects in total and $n/2$ in each sequence group. In this case, if $\delta = \mu_T - \mu_R$, $\text{Var}(\hat{\delta}) = 2\sigma_W^2/n$. The $t-$statistics for testing each of H_{01} and H_{02} are, respectively, $t_L = (\hat{\delta} + \Delta)/\sqrt{2\hat{\sigma}_W^2/n}$ and $t_U = (\hat{\delta} - \Delta)/\sqrt{2\hat{\sigma}_W^2/n}$, and each has $(n-2)$ degrees of freedom. Both hypotheses are rejected if $t_L \leq -t_{1-\alpha,n-2}$ and $t_U \geq t_{1-\alpha,n-2}$, where $t_{1-\alpha,n-2}$ is the upper $(1-\alpha)$ percentile of the central t-distribution on $n-2$ degrees of freedom. To calculate power we need to consider the joint distribution of t_L and t_U on the alternative hypothesis set of values for δ, where $\delta \neq 0$. This joint distribution is the bivariate noncentral $t-$distribution, with noncentrality parameters nc_L and nc_U, where

$$nc_L = \frac{\mu_T - \mu_R + \Delta}{\sqrt{2\sigma_W^2/n}} \quad \text{and} \quad nc_U = \frac{\mu_T - \mu_R - \Delta}{\sqrt{2\sigma_W^2/n}}. \tag{3.10}$$

The power of the TOST procedure is:

$$power(\delta, \alpha, n, \sigma_W) = Pr(t_L \leq -t_{1-\alpha,n-2} \text{ and } t_U \geq t_{1-\alpha,n-2}) \tag{3.11}$$

and is a function of δ for given n, α and σ_W. The function can be calculated by making use of the results given by [321], as done by [341]. Figure 3.9 shows some plots of the power function for two different values of σ_W and a selection of values of n. In all cases $\alpha = 0.05$. The values of the power were calculated using SAS programs kindly supplied by Dr. Klem Phillips. The values of δ and σ_W have been expressed relative to μ_R, i.e., $\delta = 100(\mu_T - \mu_R)/\mu_R$ and σ_W is a fraction of μ_R. So, for example, a value of 0.1 implies $\sigma_W = 0.1\mu_R$. The curves have been drawn for $n = 10, 20, 40,$ and 60. Grid lines have been added in the horizontal direction to indicate where the power is 0.05, 0.50, 0.80, and 0.90, respectively, from the bottom up. The vertical grid line on the right of each plot indicates the value of $\Delta = \ln(1.25)$. We can see that the power is a maximum when $\delta = 0$ and declines as δ increases. The power is 0.05 at $\delta = \Delta$, by definition, and consequently we see that all the curves pass through the point $(\Delta, 0.05)$.

Using the SAS program supplied by Dr. Phillips we have, by trial and error, calculated the samples sizes needed to achieve a power of 80% or 90% for a range of values of $\delta = 100(\mu_T - \mu_R)/\mu_R$ and σ_W. As before, the

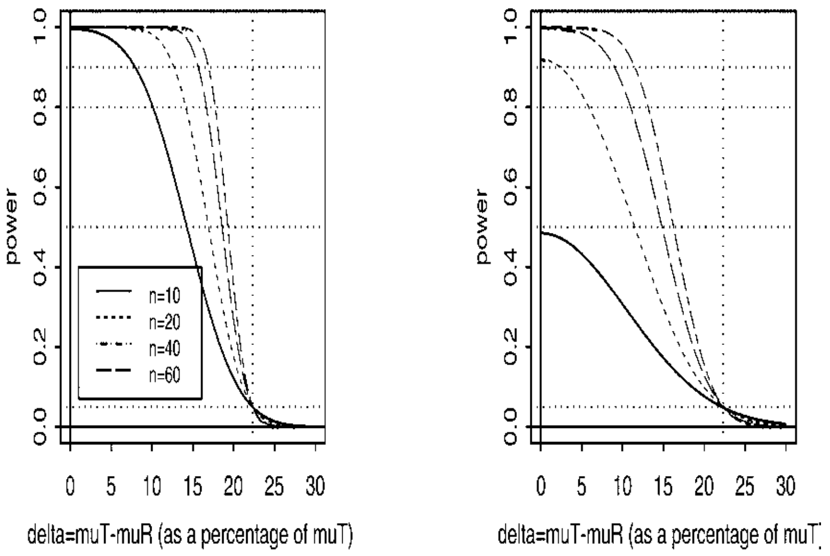

Figure 3.10 *Examples of Power Curves for the TOST Procedure. Left panel:* $\sigma_W = 0.1$*. Right panel:* $\sigma_W = 0.2$*.*

Table 3.11 *Samples Sizes for a 2×2 Cross-over Trial to Detect ABE*

f(for σ_W)	δ	80% Power	90% Power
0.1	0	6 (86.32)	8 (97.58)
	5	8 (91.77)	8 (91.77)
	10	10 (80.67)	14 (92.25)
	15	26 (82.04)	34 (90.43)
0.2	0	16 (82.39)	20 (91.92)
	5	20 (83.22)	26 (91.64)
	10	36 (81.99)	48 (90.77)
	15	48 (80.06)	130 (90.11)

value of σ_W is expressed as a fraction of μ_R, i.e., $\sigma_W = f * \mu_R$, for $f = 0.1$ and 0.2. These are given in Table 3.11, where, for convenience $\mu_R = 1$ and we have also included in parentheses the actual value of the power given be the calculations. A simple formula that gives an approximation to the power of the TOST procedure is given in equation (3.12). This uses the univariate noncentral t-distribution instead of the bivariate noncentral t-distribution. $CDF(x, df, nc) = P(t \leq x)$, where t has the noncentral t-distribution on df degrees of freedom and noncentrality parameter nc. Direct calculation shows this is a good approximation. For example, as a comparison with two of the values in Table 3.11, the formula gives power values of 81.96 and 90.75 for $\delta = 10$ and $f = 0.20$ for $n = 36$ and 48, respectively.

$$1 - \beta = CDF(t_{1-\alpha,n-2}, df, nc_L) - CDF(t_{1-\alpha,n-2}, df, nc_U). \qquad (3.12)$$

3.7 Example Where Test and Reference Are Not ABE

You may recall in the Introduction to Chapter 2 that Lenny and Denny, the ClinPharm physician and scientist, were concerned about a particular set of data from a BE trial. These data are given in Table 3.12.

Table 3.12: Example 3.2

	Sequence RT			
	AUC Period		Cmax Period	
Subject	1	2	1	2
1	58.160	79.340	2.589	2.827
3	69.680	85.590	2.480	4.407
5	121.840	.	5.319	.
8	208.330	377.150	9.634	11.808
10	17.220	14.230	1.855	1.121
11	1407.900	750.790	13.615	6.877
13	20.810	21.270	1.210	1.055
15	.	8.670	0.995	1.084
18	203.220	269.400	7.496	9.618
20	386.930	412.420	16.106	12.536
21	47.960	33.890	2.679	2.129
24	22.700	32.590	1.727	1.853
26	44.020	72.360	3.156	4.546
27	285.780	423.050	8.422	11.167
31	40.600	20.330	1.900	1.247
32	19.430	17.750	1.185	0.910
36	1048.600	1160.530	18.976	17.374
37	107.660	82.700	5.031	6.024
39	469.730	928.050	6.962	14.829
R=Reference, T=Test				

Table 3.12: Example 3.2

	Sequence RT			
	AUC Period		Cmax Period	
Subject	1	2	1	2
43	14.950	20.090	0.987	2.278
44	28.570	28.470	1.105	1.773
45	379.900	411.720	12.615	13.810
47	126.090	46.880	6.977	2.339
50	75.430	106.430	4.925	4.771
	Sequence TR			
Subject	1	2	1	2
2	150.120	142.290	5.145	3.216
4	36.950	5.000	2.442	0.498
6	24.530	26.050	1.442	2.728
7	22.110	34.640	2.007	3.309
9	703.830	476.560	15.133	11.155
12	217.060	176.020	9.433	8.446
14	40.750	152.400	1.787	6.231
16	52.760	51.570	3.570	2.445
17	101.520	23.490	4.476	1.255
19	37.140	30.540	2.169	2.613
22	143.450	42.690	5.182	3.031
23	29.800	29.550	1.714	1.804
25	63.030	92.940	3.201	5.645
28	.	.	0.531	0.891
29	56.700	21.030	2.203	1.514
30	61.180	66.410	3.617	2.130
33	1376.020	1200.280	27.312	22.068
34	115.330	135.550	4.688	7.358
38	17.340	40.350	1.072	2.150
40	62.230	64.920	3.025	3.041
41	48.990	61.740	2.706	2.808
42	53.180	17.510	3.240	1.702
46	.	.	1.680	.
48	98.030	236.170	3.434	7.378
49	1070.980	1016.520	21.517	20.116
R=Reference, T=Test				

We can see that data has been collected on 49 subjects. Some subjects have missing data points (see Subjects 28 and 46, for example). Before modelling the data we should understand why these subjects did not

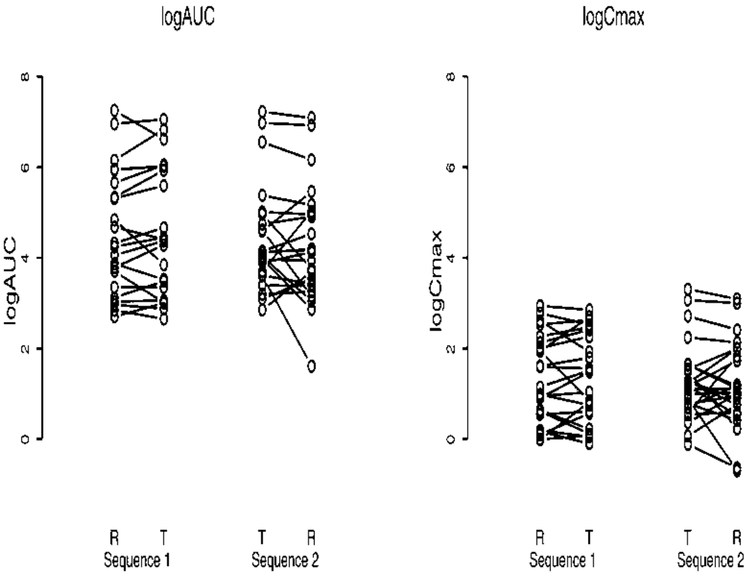

Figure 3.11 *Example 3.2: Subject Profiles Plot*

produce PK data. In the case of Subject 28, the PK concentrations were too low to produce a quality AUC value, and Subject 46 similarly did not get much drug on board after taking each dose. Subject 35 decided not to participate in the trial, and thus had no data. These things happen in bioequivalence trials and will be discussed in more detail in Chapter 5.

The subject profiles plots for these data are given in Figure 3.11, where we have included only those subjects that had two data points for either logAUC or logCmax. We can see clearly that for one subject (4 in sequence TR) there is a dramatic change from T to R for both logAUC and logCmax. This subject had particularly low AUC and Cmax values in Period 1, which though unusual were quite genuine. We will therefore leave the data for this subject in the set to be analyzed. The paired-agreement plots are given in Figure 3.12 and do not suggest that there is a significant difference between T and R, although there may be a period difference. Before we can continue to fit a linear model to the (log-transformed) data, we must decide what to do with the data from those subjects who did not provide a value in both periods. The most precise comparison of T and R is based on the difference of two values on the same subject. Such a comparison is not possible for those subjects with only a single value. If however, there are two such subjects

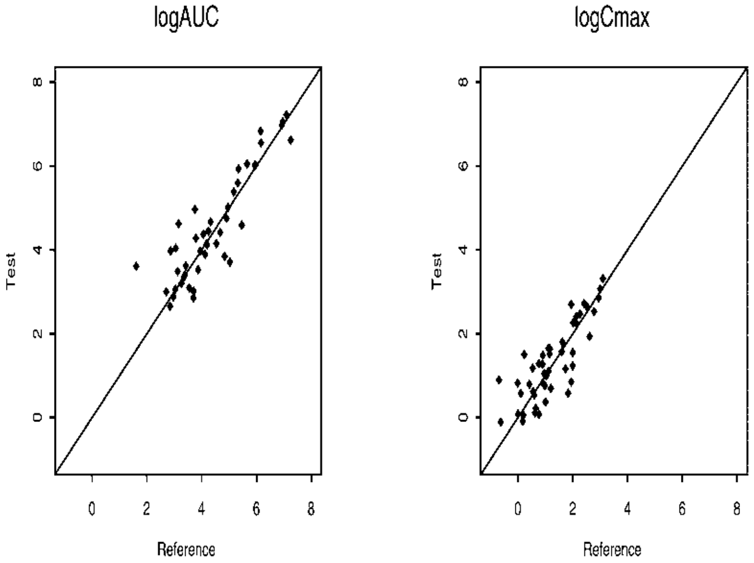

Figure 3.12 *Example 3.2: Paired-agreement Plots*

and one has a value only on T and the other has a value only on R, then a between-subject comparison of T and R is possible by taking the difference of these two single values. However, the precision of such a comparison will be low because the between-subject variation, as we can see from Figure 3.11, is much higher than the within-subject variability. Because we have assumed the subject effects ξ_{ik} are random variables, these between-subject comparisons can be recovered in the analysis if we fit what is known as a mixed model. A full explanation of mixed models is beyond the scope of the present chapter and so we will proceed to analyse the subset of data from those subjects that provided values on both T and R. We will consider mixed models in more detail in Chapter 5. However, the recovery of between-subject information on the comparison of T and R is unlikely to make much difference to the results, and so nothing of significance will be lost by ignoring the data on those subjects that provided only a single value. To justify this assertion, we will also report the results of fitting the mixed model to the complete data set, but as already mentioned a full explanation of how this was done will have to wait until Chapter 5.

The groups-by-periods means for Example 3.2 are given in Table 3.13 and plotted in Figure 3.13. The pattern is similar for both logAUC and logCmax, although there is a larger difference between T and R in the

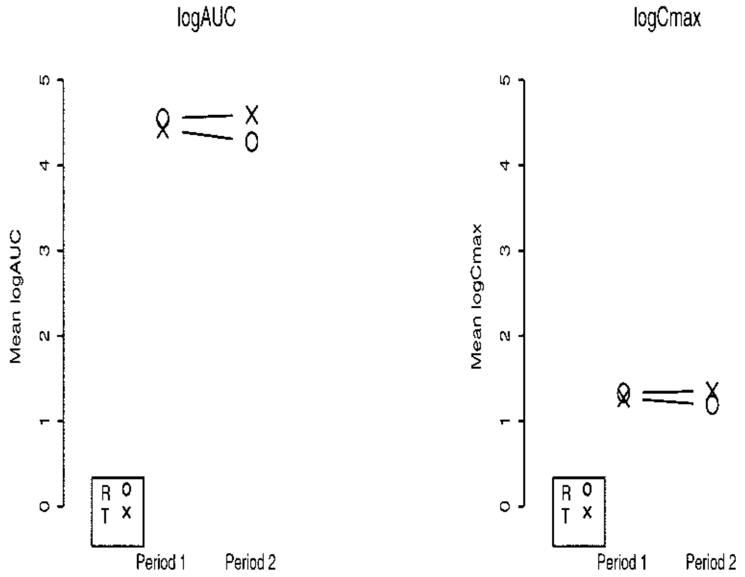

Figure 3.13 *Example 3.2: Groups-by-Periods Plot*

second period for logAUC. The mean differences versus totals plot is given in Figure 3.14. For logAUC there is a noticeable vertical separation of the centroids suggesting a possible lack of ABE.

The results of applying the TOST procedure are given in Table 3.14. We can see that the upper limit of the 90% confidence interval for logAUC (and of course AUC) is above the upper boundary for ABE. Therefore, even though ABE is not contradicted when the logCmax data are used, T and R are judged to have failed the FDA criteria for ABE. The fitted normal densities corresponding to the TOST results are given in Figure 3.15. We can see a large part of the density for logAUC extends to the right of the ABE limit and is consistent with the lack of ABE found by the TOST procedure. As with out previous example, we look at the normal probability plots to check on our assumptions. These are given in Figure 3.16. There is very strong evidence that the studentized residuals from the model for logAUC are not normally distributed. The p-value from the Shapiro-Wilk test is 0.012 for logAUC and 0.407 for logCmax. This confirms the visual indication that studentized residuals for logAUC are not normally distributed. The largest studentized residual for logAUC is 3.341, from subject 4. This is unusually large in a sample of size 45 from the standard normal distribution and the data value corresponding to this residual is an 'outlier' - i.e., a value that is

Table 3.13 *Example 3.2: Groups-by-Periods Means (sample size in brackets)*

logAUC			
Group	Period 1	Period 2	Mean
1(RT)	$\bar{y}_{11.} = 4.55(22)$	$\bar{y}_{12.} = 4.60(22)$	$\bar{y}_{1..} = 4.57$
2(TR)	$\bar{y}_{21.} = 4.43(22)$	$\bar{y}_{22.} = 4.28(22)$	$\bar{y}_{2..} = 4.35$
Mean	$\bar{y}_{.1.} = 4.49$	$\bar{y}_{.2.} = 4.43$	$\bar{y}_{...} = 4.46$
logCmax			
1(RT)	$\bar{y}_{11.} = 1.33(23)$	$\bar{y}_{12.} = 1.36(23)$	$\bar{y}_{1..} = 1.34$
2(TR)	$\bar{y}_{21.} = 1.27(23)$	$\bar{y}_{22.} = 1.19(23)$	$\bar{y}_{2..} = 1.23$
Mean	$\bar{y}_{.1.} = 1.30$	$\bar{y}_{.2.} = 1.27$	$\bar{y}_{...} = 1.29$

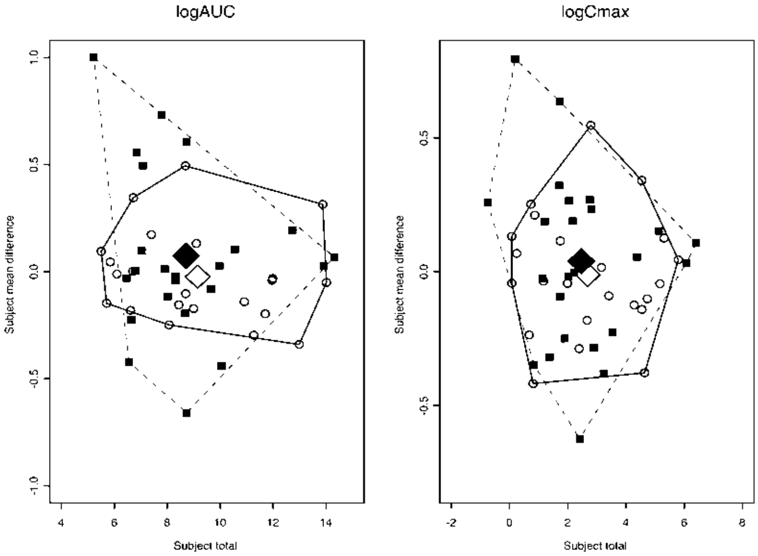

Figure 3.14 *Example 3.2: Mean Differences versus Totals Plot*

Table 3.14 *Example 3.2: TOST Procedure Results – Log Scale*

Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% Confidence Interval
logAUC (45 subjects)	0.0970	(-0.0610, 0.2550)
logCmax (47 subjects)	0.0508	(-0.0871, 0.1887)
Endpoint	**$\exp(\hat{\mu}_T - \hat{\mu}_R)$**	**90% Confidence Interval**
AUC (45 subjects)	1.10	(0.94, 1.29)
Cmax (47 subjects)	1.05	(0.92, 1.21)

unusual relative to the fitted model. As already noted, this is a subject with a very large drop in value for both logAUC and logCmax over the two periods. Subject 14 has a large increase in both logAUC and logCmax between the periods.

An alternative analysis that does not depend on the assumption that the data are normally distributed is available and we will illustrate this (nonparametric analysis) in the next section. It should be noted, however, that regulatory approval may not be obtained if nonparametric methods are used.

To end this chapter we report the results that are obtained by fitting a mixed model to the complete data set. These are displayed in Table 3.15 and lead to the same conclusions as were obtained from the data using only those subjects with values in both Period 1 and Period 2.

Table 3.15 *Example 3.2: TOST Procedure Results (all subjects) – Log Scale*

Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% Confidence Interval
logAUC	0.0940	(-0.0678, 0.2482)
logCmax	0.0468	(-0.0907, 0.1843)
Endpoint	**$\exp(\hat{\mu}_T - \hat{\mu}_R)$**	**90% Confidence Interval**
AUC	1.09	(0.93, 1.28)
Cmax	1.05	(0.91, 1.20)

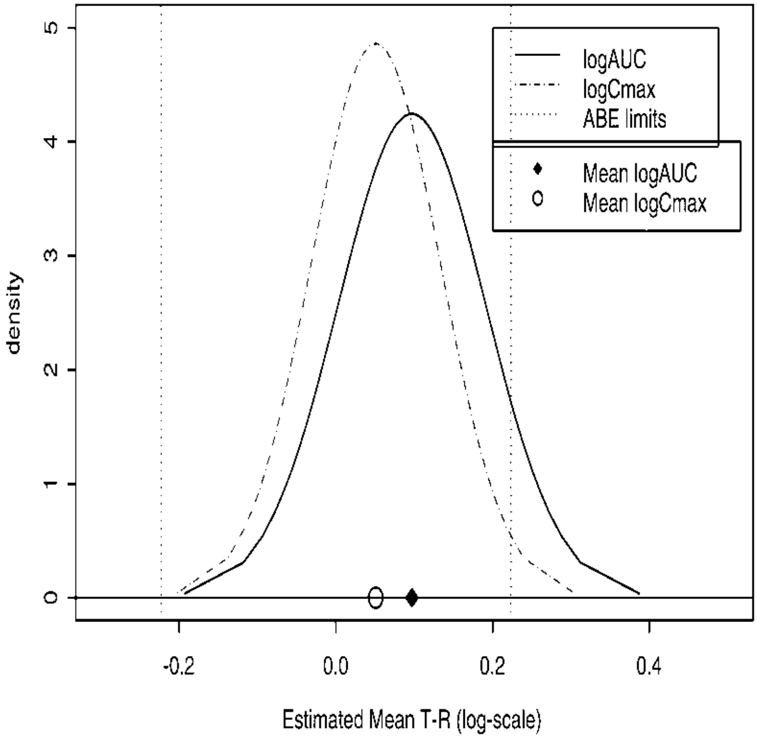

Figure 3.15 *Example 3.2: Fitted Normal Densities for* $\hat{\mu}_T - \hat{\mu}_R$

3.8 Nonparametric Analysis

On occasion, in bioequivalence or relative bioavailability studies, there may be a need to analyse unusual endpoints beyond those usually assessed such as logAUC and logCmax. Examples of such endpoints are:

1. The ratio Cmax/AUC [102] - an alternative measure of the rate of exposure
2. Partial AUC [105] - an alternative measure of absorption
3. λ (see Equation 1.1, Chapter 1) - a measure of excretion.

Endpoints like these are difficult to assess using models like those introduced up to now as they are unlikely to be normally distributed. For example, the ratio Cmax/AUC is the ratio of two log-normally distributed variables. Even though it may be possible to derive approximate or exact formulae for the distributions of such endpoints, it is unclear how this would directly benefit the sponsors of such studies or patients. These

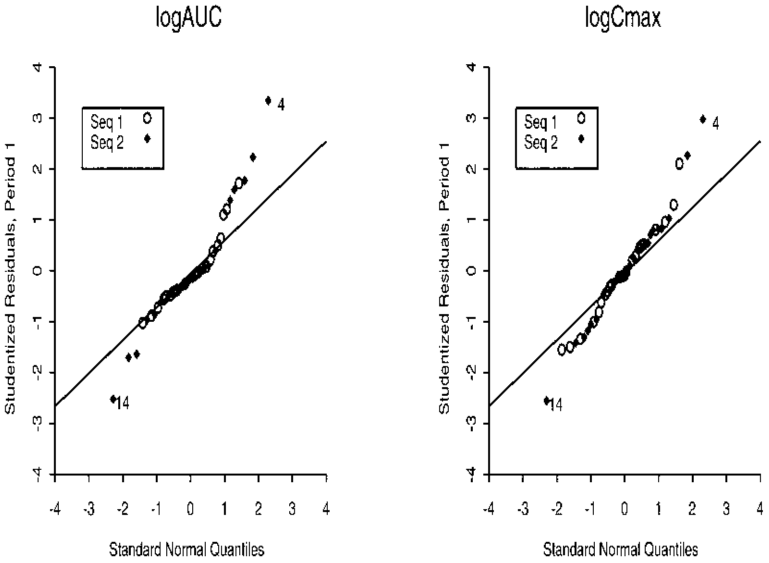

Figure 3.16 *Example 3.2: Normal Probability Plots*

endpoints are currently viewed as supportive only, and exact quantification of their Type 1 or 2 error rates are not of immediate concern in a regulatory filing.

However, there are statistical procedures available to analyse such (non-normal) data. These procedures are termed ‘nonparametric’ in that they do not assume a particular parametric form (e.g., normal or lognormal) for the endpoint of interest. The nonparametric analysis for the 2 × 2 cross-over was first described by [253] and later illustrated by [81] in the context of evaluating bioavailability data. An excellent review of nonparametric methods for the analysis of cross-over trials is given by [434]. See also [414] and [198]. For a more extensive coverage of the methods covered in this section see [420].

Such nonparametric analyses should only be utilised when (i) an endpoint is grossly non-normal or (ii) cannot be transformed to an endpoint that is normally distributed or (iii) sample size does not permit the application of the central-limit theorem. Such an endpoint is Tmax, which is often used to support parametric analysis findings from the analysis of logAUC and logCmax. Discussion is included here for completeness.

Obviously, the interpretation of nonparametric analysis from a regulatory perspective is overshadowed by global regulatory recommendations (see Chapter 2) on provision of adequate sample size to support parametric interpretation. Such nonparametric techniques are generally of interest to sponsors only when small sample sizes are employed and, even then, only when analysing Tmax or unusual endpoints. If there is evidence that the log-transformed data from an ABE trial are such that it would be unreasonable to assume that they are normally distributed, then the usual two one-sided t-tests (as used in the TOST procedure), can be replaced by Wilcoxon rank-sum tests, or equivalently by Mann-Whitney U-tests. As with normal data these nonparametric tests are based on within-subject differences.

To illustrate the nonparametric tests we will use the Tmax values recorded on the subjects in Example 3.1. These are given in Table 3.16.

In order to make a comparison between the parametric and nonparametric procedures we will first analyse the logTmax values as if they were normally distributed (i.e., on the assumption the Tmax values are log-normally distributed). Although there are no regulatory guidelines on what must be done to determine if T and R are ABE, we will apply the same regulatory hurdles that apply to logAUC and logCmax. In other words we will apply the usual TOST procedure to logTmax. The results of doing this, along with the back-transformed values, are given in Table 3.17. Applying the familiar regulatory guidelines, T and R cannot be deemed to be ABE, as the upper 90% confidence limit for $\mu_T - \mu_R$, on the logTmax scale, exceeds 0.2231 (and, of course, the upper 90% confidence limit on the Tmax scale exceeds 1.25).

However, we need to check that the assumption that the residuals from our usual linear model (3.1) for logTmax are normally distributed is reasonable. Figure 3.17 displays the histogram of the studentized residuals and a normal probability plot. The studentized residuals look like they can be assumed to be normally distributed. The p-value for the Shapiro-Wilk test for normality is 0.7044, which also gives some assurance that the studentized residuals have a normal distribution. However, a closer inspection of the normal probability plot in Figure 3.10 reveals horizontal bands of residuals; a feature most unlikely to occur if the residuals were normally distributed. Also, of course, the nature of the Tmax variable itself indicates that logTmax will not be normally distributed. The concentrations are only taken at a set of predetermined times, and so Tmax is an inherently discrete random variable.

A further warning sign is that when Model (3.1) was fitted using PROC MIXED, the estimate of $\hat{\sigma}^2_B$ (not shown) was zero, indicating some instability in the REML fitting procedure for these data. The values in Table 3.18 were therefore calculated using the results of fitting Model (3.1) under the assumption that the subject parameters were fixed rather than random effects. Of course, for a complete data set like that in Table 3.16, with two values of Tmax for every subject, we should get the same

TOST results irrespective of whether the subject parameters are fixed or random. The fact that we do not is another indication that a more robust analysis procedure should be used for these data.

When data are log-normal or normal in distribution, it is known that, in most cases, the probability of a Type 2 error is increased when using a nonparametric procedure relative to the parametric procedures discussed in earlier sections [193].

Table 3.16: Example 3.1: Tmax

Subject	Sequence	Period	Formulation	Tmax	LogTmax
1	RT	1	R	0.50	-0.693
1	RT	2	T	0.50	-0.693
4	RT	1	R	0.50	-0.693
4	RT	2	T	1.00	0.000
5	RT	1	R	1.50	0.405
5	RT	2	T	0.25	-1.386
8	RT	1	R	1.00	0.000
8	RT	2	T	0.50	-0.693
9	RT	1	R	0.25	-1.386
9	RT	2	T	1.50	0.405
11	RT	1	R	0.50	-0.693
11	RT	2	T	1.00	0.000
16	RT	1	R	1.50	0.405
16	RT	2	T	2.00	0.693
17	RT	1	R	1.50	0.405
17	RT	2	T	1.00	0.000
19	RT	1	R	1.50	0.405
19	RT	2	T	0.50	-0.693
21	RT	1	R	0.50	-0.693
21	RT	2	T	0.50	-0.693
24	RT	1	R	1.00	0.000
24	RT	2	T	0.50	-0.693
25	RT	1	R	1.00	0.000
25	RT	2	T	0.25	-1.386
28	RT	1	R	1.00	0.000
28	RT	2	T	1.50	0.405
29	RT	1	R	1.00	0.000
29	RT	2	T	1.50	0.405
31	RT	1	R	0.50	-0.693
31	RT	2	T	1.50	0.405
34	RT	1	R	0.50	-0.693

R=Reference, T=Test

Table 3.16: Example 3.1: Tmax

Subject	Sequence	Period	Formulation	Tmax	LogTmax
34	RT	2	T	1.50	0.405
36	RT	1	R	0.50	-0.693
36	RT	2	T	1.00	0.000
2	TR	1	T	1.00	0.000
2	TR	2	R	1.00	0.000
3	TR	1	T	0.50	-0.693
3	TR	2	R	0.50	-0.693
6	TR	1	T	1.00	0.000
6	TR	2	R	0.50	-0.693
7	TR	1	T	1.00	0.000
7	TR	2	R	0.25	-1.386
10	TR	1	T	1.50	0.405
10	TR	2	R	1.00	0.000
12	TR	1	T	1.00	0.000
12	TR	2	R	1.00	0.000
15	TR	1	T	0.50	-0.693
15	TR	2	R	1.50	0.405
18	TR	1	T	1.00	0.000
18	TR	2	R	0.50	-0.693
20	TR	1	T	1.00	0.000
20	TR	2	R	0.50	-0.693
22	TR	1	T	2.00	0.693
22	TR	2	R	4.02	1.391
23	TR	1	T	0.50	-0.693
23	TR	2	R	0.50	-0.693
26	TR	1	T	0.50	-0.693
26	TR	2	R	0.25	-1.386
27	TR	1	T	0.50	-0.693
27	TR	2	R	1.00	0.000
30	TR	1	T	0.50	-0.693
30	TR	2	R	1.00	0.000
35	TR	1	T	0.50	-0.693
35	TR	2	R	1.00	0.000
R=Reference, T=Test					

To derive the equivalent of the TOST procedure based on a non-parametric approach we use the Hodges-Lehmann point estimate and confidence interval for $\mu_T - \mu_R$ [207]. These can be calculated using tables, by asymptotic approximation or from software for exact testing

Table 3.17 *Example 3.1: TOST Procedure Results for Tmax*

Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% Confidence Interval
logTmax	0.0553	(-0.2021, 0.3126)
Tmax	1.0569	(0.8170, 1.3670)

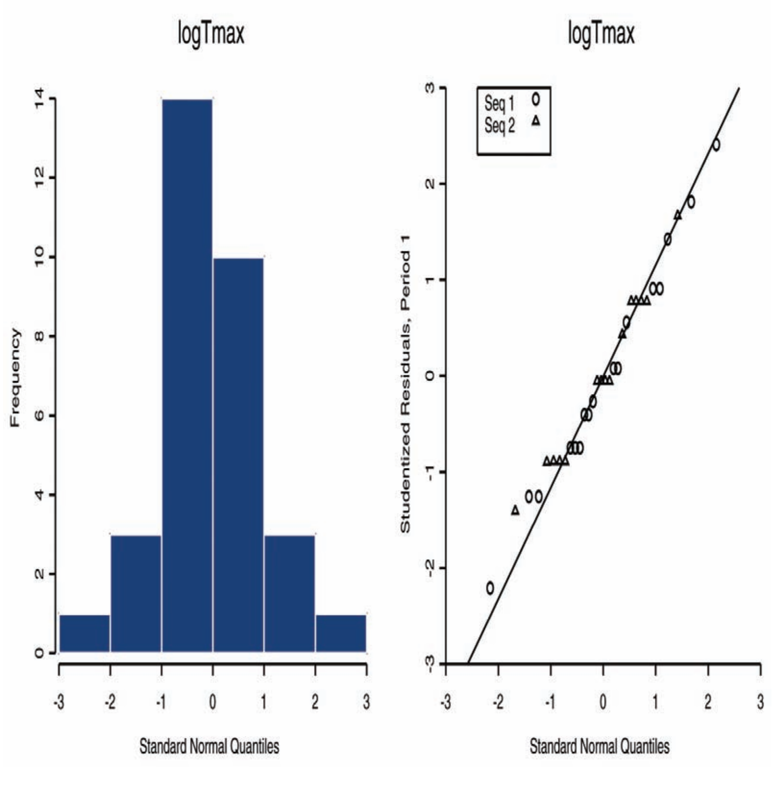

Figure 3.17 *Example 3.1: Studentized Residuals for Tmax*

such as StatXact [86]. Here we will illustrate the approach that uses the asymptotic approximation.

It will be recalled that the estimate of $\delta = \mu_T - \mu_R$ was obtained previously by comparing the mean period difference from sequence Group

2 with the mean period difference from sequence Group 1:

$$\hat{\delta} = \hat{\mu}_T - \hat{\mu}_R = \frac{1}{2}([\bar{y}_{21.} - \bar{y}_{22.}] - [\bar{y}_{11.} - \bar{y}_{12.}]).$$

The robust estimate of δ is based on similar reasoning, but uses the median rather than the mean.

In order to construct a robust equivalent of the 90% confidence interval used in the TOST procedure, we first calculate for each subject the difference between the logTmax values in Periods 1 and 2, (i.e., $y_{i1k} - y_{i2k}$, for $i = 1, 2$ and $k = 1, 2, \ldots, n_i$).

Let us label the period differences, $y_{11k} - y_{12k}$, in sequence Group 1 as X_i, $i = 1, 2, \ldots, n_1$ and the differences, $y_{21k} - y_{22k}$ in sequence Group 2 as Y_j, $j = 1, 2, \ldots, n_2$. In Example 3.1, $n_1 = 17$ and $n_2 = 15$.

To calculate the point estimate we first form the $n_1 \times n_2$ differences $Y_j - X_i$, for $i = 1, 2, \ldots, n_1$ and $j = 1, 2, \ldots, n_2$. The point estimate $\hat{\delta}$ is then half the value of the median of these differences. To obtain the median, the differences are ordered from smallest to largest. To save space, we do not give the list of these ordered differences here. If $n_1 \times n_2$ is odd and equals $2p+1$, say, the median is the $(p+1)$th ordered difference. If $n_1 \times n_2$ is even and equals $2p$, say, the median is the average of the pth and $(p+1)$th ordered differences. For Example 3.1, $n_1 n_2 = 255$ and therefore median is the 128th ordered difference, which is 0, i.e., $\hat{\delta} = 0/2$.

To obtain a symmetric two-sided confidence interval for δ, with confidence coefficient $1 - \alpha$, we must first obtain an integer, which we will denote by C_α. To get this we use the critical values of the distribution of the Wilcoxon rank-sum test statistic [212], which can be obtained by approximation when n_1 and n_2 are large (i.e., larger than 12) or from Table A.6 of [212] when n_1 and n_2 are small. The Wilcoxon rank-sum test can be considered as a nonparametric form of the usual $t-$test for comparing two independent samples. The rank-sum test uses the ranks of the data rather than the data themselves. We will say more about this test after describing and illustrating the nonparametric form of the TOST procedure.

To obtain C_α when n_1 and n_2 are small (i.e., ≤ 12) we first obtain the value $w(\alpha/2, n_1, n_2)$ from Table A.6. This value is such that, on the null hypothesis of no difference in central location between the two samples under consideration, $\mathrm{P}[W \geq w(\alpha/2, n_1, n_2)] = \alpha/2$, where W is the rank-sum statistic. The value of C_α is then obtained by noting that $[n_2(2n_1 + n_2 + 1)/2] - C_\alpha + 1 = w(\alpha/2, n_1, n_2)$. On the null hypothesis, C_α is the largest integer such that

$$\mathrm{P}\left[\left(\frac{n_2(n_2+1)}{2} + C_\alpha\right) \leq W \leq \left(\frac{n_2(2n_1+n_2+1)}{2} - C_\alpha\right)\right] \geq 1 - \alpha,$$

where $n_1 > n_2$.

For large n_1 and n_2, the integer C_α may, according to Hollander and Wolfe, be approximated by

$$C_\alpha = \frac{n_1 n_2}{2} - z_{\alpha/2}\left[\frac{n_1 n_2(n_1 + n_2 + 1)}{12}\right]^{\frac{1}{2}},$$

where $z_{\alpha/2}$ is the upper $(1 - \alpha/2)$ point of the standard normal distribution.

The $(1 - \alpha)$ confidence interval is the $\frac{1}{2}(\delta_L, \delta_U)$, where δ_L is the C_αth ordered difference and δ_U is the $(n_1 n_2 + 1 - C_\alpha)$th ordered difference.

Taking $z_{0.05} = 1.645$, we get $C_\alpha = 84$. That is, the 90% confidence interval is obtained by taking δ_L as the 84th ordered difference and δ_U as the 172nd ordered difference. The resulting 90% confidence interval for δ is $(-0.2027, 0.3466)$. The back-transformed interval is (0.82, 1.41). These are quite similar to those obtained previously, (-0.2021, 0.3126) and (0.82, 1.37), respectively, indicating some robustness of the parametric approach when the sample sizes are relatively large.

The Wilcoxon rank-sum test assumes that the endpoint (logTmax in our case) is expressed on an interval (or metric) scale, so that the same shift on the scale has the same interpretation regardless of its location. Further assumptions made in using this test include randomisation of subjects to the groups with random sampling from the same family of distributions with differences between groups only being for location.

To calculate the test statistic, the period differences are ranked, where the ranking is done in terms of the total number of subjects, not separately for each group. Let R_i = [the sum of the ranks of group i], $i = 1, 2$. Under the null hypothesis that $\mu_T = \mu_R$,

$$E[R_1] = n_1(n_1 + n_2 + 1)/2 \ ,$$

$$E[R_2] = n_2(n_1 + n_2 + 1)/2$$

and

$$\mathrm{Var}[R_1] = \mathrm{Var}[R_2] = n_1 n_2(n_1 + n_2 + 1 - T)/12 \ ,$$

where T is a correction for ties.

If there are no ties then $T = 0$. If there are v tied sets, with t_s ties in the sth set, where $s = 1, 2, \ldots, v$, then

$$T = \frac{\sum_{s=1}^{v} t_s(t_s^2 - 1)}{[(n_1 + n_2)(n_1 + n_2 - 1)]}.$$

An asymptotic test of the null hypothesis can be based on either R_1 or R_2. For R_1 we calculate

$$z = \frac{R_1 - E[R_1]}{(\mathrm{Var}[R_1])^{\frac{1}{2}}}$$

and compare it with the standard normal distribution. Statistical software, such as `proc npar1way` in SAS will do the necessary calculations for this test and produce exact P-values for small n_1 and n_2.

In order to apply the nonparametric equivalent of the TOST procedure, we use the mean difference, (Period 1 - Period 2)/2, for each subject. In the analysis we (i) add log(1.25) to the mean differences in Group 2 and apply the Wilcoxon rank sum test and (ii) subtract log(1.25) from the mean differences in Group 2 and apply the test. The P-values from the exact and asymptotic tests are very similar (0.021 when adding and 0.120 when subtracting, for the exact test) and are not very different from the $t-$test (0.038 and 0.138). Indeed the conclusions are the same; based on logTmax T and R are not ABE.

3.9 Some Practical Issues

In some cases, multiple AUC endpoints are derived in bioequivalence data sets. In general, if half-life ($T_{\frac{1}{2}}$, see Chapter 1) is estimable, it will be used to calculate AUC(0-∞) as described in Equation (1.1). However, if insufficient concentration data are captured during elimination of the drug, half-life may not be estimable, and therefore AUC(0-t) will be used in statistical evaluation. Recall that in this context t denotes the last quantifiable concentration during the period in which samples are captured. FDA guidance [135] notes that if over 20% of the value of AUC(0-∞) is attributable to calculation with $T_{\frac{1}{2}}$ then AUC(0-t) should be used in bioequivalence assessment.

The 't' in AUC(0-t) may differ across periods for any given subject. For example, in Example 3.1, subject 15's last quantifiable concentration occurred in the first period at 12 hours and in the second at approximately 16 hours post dose. In data sets where marked differences between the last quantifiable time are present between periods and half-life is inestimable, it may be preferable to consider an endpoint like AUC(0-t'). Here, t' denotes the last quantifiable concentration time in common across periods for a given subject.

The decision about which AUC endpoint is primary and which will provide supportive information should be made prior to analysis to prevent the introduction of bias into interpretation of the data [388]. In the authors' experience, AUC(0-∞) is most often used, with AUC(0-t) used in those cases where half-life is not estimable for a large number of subjects. It is unusual for AUC(0-t') to be used as most BE studies are designed to ensure sufficient samples are taken during elimination to ensure half-life is estimable. FDA guidance [135] recommends that both AUC(0-t) and AUC(0-∞) be provided in submissions.

In cases where multiple 'peaks' in blood concentration are observed, it is common practice for the first [131] to be chosen as Cmax, with the corresponding time relative to dose being Tmax. The value of Tmax is highly dependent on the choice of sampling times. Its use in bioequivalence studies is that of an endpoint providing supportive information. Some nations [57]-[58] require that Tmax be analyzed as if it were normally distributed.

CHAPTER 4

BE Studies with More Than Two Periods

Introduction

Denny walked into my office one day after the reports for Example 3.2 came out looking like he had been run over by a bus and dragged over hot coals. He had been (figuratively) when he reviewed the findings with senior management. They obviously did not like the implications for getting together a marketable formulation in time for filing with the FDA.

Nobody ever comes to see you when you release findings they like. That annoyed me when I first started on the job, but after a while I realized it gave one more time to enjoy the moment.

Do take time out to enjoy the good moments on the job. Given the success rate of drugs in clinical development (see Chapter 1), statisticians should expect to be the bearer of bad news on the majority of occasions in their working life. This is ok if you are in an organization that recognizes that failure is far more common in drug development than success, but if you are not, grow a thick skin about such matters, or think about changing jobs. Be careful not to get cynical, though. It is an easy trap to fall into and causes one to not enjoy anything (because you always think about the bad thing that is probably right around the corner and guard against keeping your hopes up). Probabilistically speaking, there will be good moments on the job, and one should maintain one's equanimity so that one can enjoy them.

The question Denny posed to me was simple on the surface - can we explore these data to see if there was any possibility of a follow-up bioequivalence trial being successful?

Note the careful use of the word 'we'. When a clinician uses 'we' with a statistician, it is the royal 'we' which can be usually translated as meaning 'you'.

I told him that yes, ***I*** *could, but given the findings of Example 3.2, my intuition told me that it was going to be pretty unlikely and that he had better prepare his folks for that message. I would run some programs and get back to him with a quantitative assessment next week. He wanted it sooner, but I told him no.*

I got through to Denny on three of four points here (which is pretty good all things considered). He recognized that I would do the work by next week and that the success of a follow-up study was going to be low, but the idea that he should warn his folks went in one ear and out the other. Maybe clinicians like surprises - I gave up on trying to figure that one out long ago.

Statisticians should also recognize one other truth in drug development which people tend not to mention when they are hiring you. One would think that statisticians would recognize this fact (i.e., we are trained to count), but it seems like it gets by a lot of us. The fact is statisticians are outnumbered in drug development! There are a lot more scientists, clinicians, etc., who need our expertise than there are time or personnel to deliver it.

Hence, an option one sometimes considers as a biostatistician is to go with one's intuition and not spend the time quantifying precisely questions like that posed by Denny. We encourage people not to make the choice to opt out of applying statistical expertise in such situations. It is important to the patients who will be using such medications that we get it right. If worse comes to worse, we recommend taking the time to train the scientists and clinicians to do such work themselves.

4.1 Background

Although the RT/TR design is often the design of choice when testing for ABE, there are situations where a design with more than two periods is needed. These include

- The drugs to be compared are highly variable;
- Carry-over effects cannot be entirely ruled out due to long half-life, poor metabolism, or other factors inhibiting elimination.

By definition, a drug that is highly variable has a large within-subject variance σ^2_W (for logAUC or logCmax). Typically this is taken to mean that $\sigma^2_W \geq 0.09$ for R. Consequently, the estimate of $\mu_T - \mu_R$ will also have a large variance unless many subjects are enrolled. As large ABE trials are unattractive for ethical, statistical, and financial reasons, a better alternative is needed. If more than two periods can be used then suitable alternative designs are available. The regulatory guidance recommends using four-period, two-sequence designs such as RTRT/TRTR when highly variable drugs are compared. Such four-period designs are also needed when individual bioequivalence is considered, as we shall see in Chapter 6.

However, if the time available for the trial does not permit four periods to be used then a three-period design, with sequences such as RTT/TRR, can be used. We will review and compare these designs in the next section. In Section 4.3 we will review and compare the various four-period designs. In each of these sections we will illustrate the analysis of data and give an example of at least one such trial.

ABE trials are not confined to comparing Test and Reference. Sometimes two alternative versions of Test or Reference are included, leading to the need for designs for three or four formulations. For example, in a confirmatory trial a 300 mg Test tablet was given either (i) as 3×100 mg tablets or (ii) as a 200 mg tablet plus a 100 mg tablet. This was because in the early stages of the confirmatory trial only the 3×100 mg version was available. Later, a 200 mg tablet became available. The commercial formulation of the drug was to be a single 300 mg tablet, and this had to be shown to be ABE to the versions used in the confirmatory trial. A trial with four formulations might arise when both a high and a low dose of Test are to be compared to a high and low dose of Reference. Examples of both of these types of design will be given in Section 4.6. The data sets for each example are given in Section 4.8.

4.2 Three-period Designs

As already discussed, the need for extra periods usually arises when the drugs being compared are highly variable. Adding an extra period to the RT/TR design is a simple way of increasing the number of responses collected from each subject. In addition, as we shall see, a suitably chosen three-period design can give some protection against the occurrence of (unequal) carry-over effects of T and R.

Here we will only consider designs with two-sequences and the only three choices worth considering (see [237], Ch. 3) are the following, where the rows are the sequences and the columns are the periods. We assume that there are $n/2$ subjects assigned to each sequence:

1.	R	T	T	2.	R	T	R	3.	R	R	T
	T	R	R		T	R	T		T	T	R

The question now arises as to which one of these should be used. If there are no (differential) carry-over effects then the three designs are equivalent and any one may be used; the regulatory guidelines express a preference for the RTR/TRT design. However, if differential carry-over effects (i.e., $\lambda_T \neq \lambda_R$) cannot be ruled out, then the first design RTT/TRR is to be preferred, as we will shortly demonstrate.

However, before doing this let us consider the estimation of $\delta = \mu_T - \mu_R$. As an illustration we will do this for the first design given above, RTT/TRR.

Let $\bar{y}_{ij.}$ denote the mean of the response (logAUC or logCmax) in period j of sequence group i, where, as already stated, there are $n/2$ subjects in each group. Using an obvious extension of the notation used

Table 4.1 *The Expectations of $\bar{y}_{ij.}$ for Design 1*

Group	Period		
	1	2	3
1 RTT	$\gamma_1 + \pi_1 + \tau_R$	$\gamma_1 + \pi_2 + \tau_T + \lambda_R$	$\gamma_1 + \pi_3 + \tau_T + \lambda_T$
2 TRR	$\gamma_2 + \pi_1 + \tau_T$	$\gamma_2 + \pi_2 + \tau_R + \lambda_T$	$\gamma_2 + \pi_3 + \tau_R + \lambda_R$

in Chapter 3, the expectations of the six group-by-period means are given in Table 4.1.

For our illustrative design, the estimator of δ is:

$$\hat{\delta} = (-2\bar{y}_{11.} + \bar{y}_{12.} + \bar{y}_{13.} + 2\bar{y}_{21.} - \bar{y}_{22.} - \bar{y}_{23.})/4 \qquad (4.1)$$

and $\text{Var}[\hat{\delta}] = 3\sigma_W^2/(2n)$. It is easily confirmed that this is an unbiased estimator:

$$\begin{aligned} E[-2\bar{y}_{11.} + \bar{y}_{12.} + \bar{y}_{13.}] &= -2(\gamma_1 + \pi_1 + \tau_R) + (\gamma_1 + \pi_2 + \tau_T + \lambda_R) \\ &+ (\gamma_1 + \pi_3 + \tau_T + \lambda_T) \\ &= -2\pi_1 + \pi_2 + \pi_3 - 2\tau_R + 2\tau_T + \lambda_R + \lambda_T \end{aligned}$$

and

$$E[-2\bar{y}_{21.} + \bar{y}_{22.} + \bar{y}_{23.}] = -2\pi_1 + \pi_2 + \pi_3 - 2\tau_T + 2\tau_R + \lambda_R + \lambda_T.$$

Taking the second expression away from the first leaves $4(\mu_T - \mu_R)$.

The unbiased estimator of $\lambda_T - \lambda_R$ is:

$$\widehat{\lambda_T - \lambda_R} = (-\bar{y}_{12.} + \bar{y}_{13.} + \bar{y}_{22.} - \bar{y}_{23.})/2, \qquad (4.2)$$

which again can be easily confirmed. The variance of this estimator is $2\sigma_W^2/n$.

An interesting and important property of the these two estimators is that they have a covariance of zero, which for normally distributed data implies they are independent. In other words, if we were to drop the carry-over parameters from the above model, we would get the same estimator of δ as given in (4.1).

We now return to answer the question of which design out of the three possibilities given above is to be preferred. To compare the designs it is useful to use the concept of *efficiency*, which is more fully explained in Section 4.8, the Technical Appendix. Defining $\delta = \mu_T - \mu_R$, as before, efficiency is the ratio of $\text{Var}(\hat{\delta})$ in the design under consideration to the value this variance would take in an 'ideal' design. In the ideal design the effects of subjects, periods and carry-over effects can be removed from the estimate of δ. Therefore, in the ideal design, the estimate of δ, using logAUC, for example, would simply be the difference $\bar{y}_T - \bar{y}_R$, where $\bar{y}_T(\bar{y}_R)$ is the mean of all the logAUC values of $T(R)$. If T and R each occurred r times in the design then $\text{Var}(\bar{y}_T - \bar{y}_R) = 2\sigma_W^2/r = 4\sigma_W^2/(3n)$, as $r = 3n/2$. Such a design may not exist, and its use is merely to provide a lower bound on $\text{Var}(\hat{\delta})$ that may be used as a point of reference when comparing designs. Efficiency is usually expressed as a percentage, and so a fully efficient design has a value of 100%. In the presence of differential carry-over effects, the efficiency of the first design is $(4\sigma_W^2/3n)/(3\sigma_W^2/2n) \times 100 = 88.9\%$. The efficiencies of the other designs can be calculated similarly (see Section 4.8) and are 22.2% and 66.7%, respectively. In addition, as already noted, the correlation between $\hat{\delta}$ and $\widehat{\lambda_T - \lambda_R}$ in the first design is zero, whereas in the second and third designs it is 0.87 and 0.50, respectively. In other words, the first design is not only highly efficient in the presence of differential carry-over effects, but is such that the estimator of δ is the same whether or not carry-over effects are entered into the model for the data. Consequently, there is no disadvantage in using this design even if differential carry-over effects are anticipated or cannot be avoided.

4.2.1 Examples of the Analysis of BE Trials with Three-Periods

Example 4.1

The data in Tables 4.25 and 4.26 are from a trial that used the sequences RTT and TRR. Figure 4.1 shows the corresponding subject profiles plots. The most noteworthy feature in these plots is that, although the between-subject variability is high for both metrics it is much lower for logCmax compared to logAUC. In addition, the maximum value in each period for logCmax is much lower the corresponding maximum for logAUC. There is a suggestion for Sequence 2 that the values of R are higher on average than those of T, but this feature is not so evident in Sequence 1. We can also identify a subject in Sequence 2 that only provided two logAUC values.

The group-by-period means are given in Table 4.2, where because of the missing data, we have indicated the number of subjects that provided data for each mean. These are plotted in Figure 4.2, where the lower line in each plot refers to Sequence RTT and the upper line to Sequence TRR. Despite the difference in absolute size of the logAUC and logCmax means, there is a similar pattern of formulation differences within each period for both metrics. The only other notable feature is that the means for Sequence 2 are consistently higher than the corresponding means for Sequence 1.

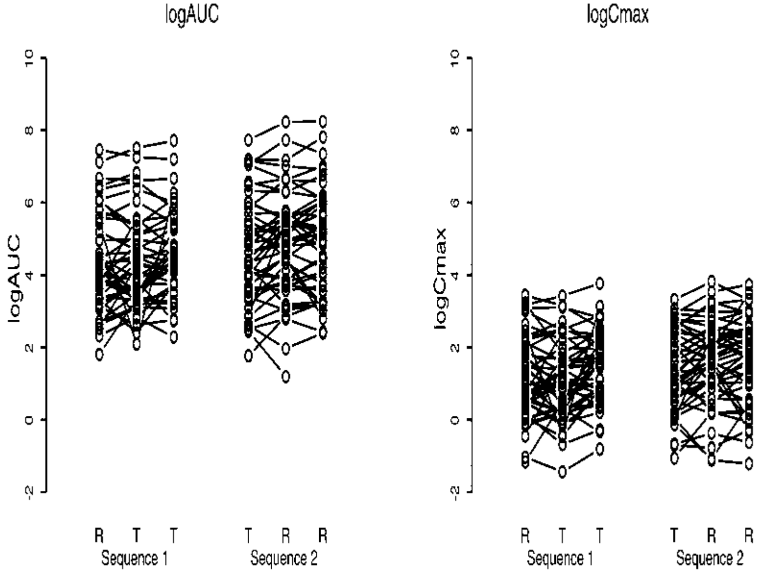

Figure 4.1 *Example 4.1: Subject Profiles Plot*

Table 4.2 *Example 4.1: Group-by-Period Means (sample size in brackets)*

		logAUC		
Group	Period 1	Period 2	Period 3	Mean
1(RTT)	$\bar{y}_{11.} = 4.35(46)$	$\bar{y}_{12.} = 4.36(45)$	$\bar{y}_{13.} = 4.60(43)$	$\bar{y}_{1..} = 4.43$
2(TRR)	$\bar{y}_{21.} = 4.66(47)$	$\bar{y}_{22.} = 4.88(47)$	$\bar{y}_{23.} = 4.92(47)$	$\bar{y}_{2..} = 4.82$
Mean	$\bar{y}_{.1.} = 4.51$	$\bar{y}_{.2.} = 4.63$	$\bar{y}_{.3.} = 4.77$	$\bar{y}_{...} = 4.63$
		logCmax		
1(RTT)	$\bar{y}_{11.} = 1.18(47)$	$\bar{y}_{12.} = 1.10(47)$	$\bar{y}_{13.} = 1.46(45)$	$\bar{y}_{1..} = 1.24$
2(TRR)	$\bar{y}_{21.} = 1.39(48)$	$\bar{y}_{22.} = 1.60(48)$	$\bar{y}_{23.} = 1.64(48)$	$\bar{y}_{2..} = 1.54$
Mean	$\bar{y}_{.1.} = 1.29$	$\bar{y}_{.2.} = 1.35$	$\bar{y}_{.3.} = 1.55$	$\bar{y}_{...} = 1.40$

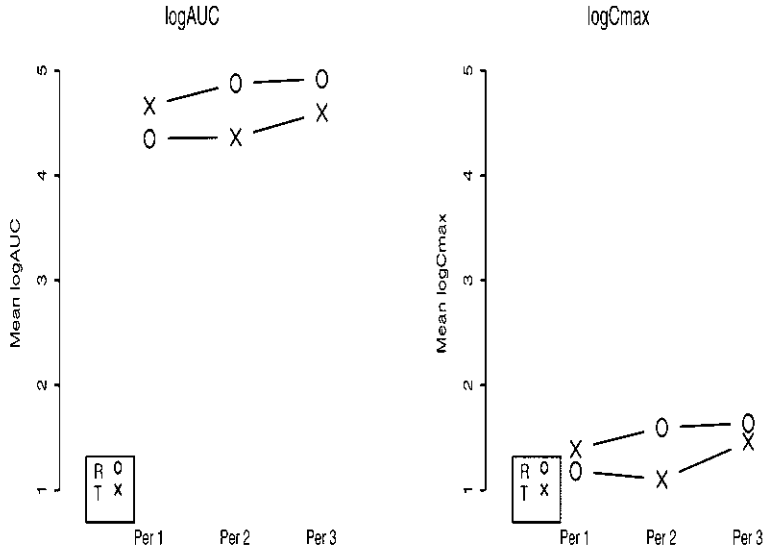

Figure 4.2 *Example 4.1: Groups-by-periods Plot*

To get a graphical impression of the similarity or otherwise of the means of R and T, we can use a version of the mean differences versus totals plot that was used in Chapter 3 for the RT/TR design. In this alternative version of the plot we replace the within-subject mean difference with a within-subject contrast for the kth subject in sequence group i: $d_{ik} = -(2y_{i1k} - y_{i2k} - y_{i3k})/4$. From Equation (4.1), we can see that $\hat{\delta} = \bar{d}_{1.} - \bar{d}_{2.}$. Instead of the subject totals, we arbitrarily, use the mean of each subject, so that we can plot the subject contrasts against the subject means. If the contrasts are plotted on the vertical axis, any separation of the groups along this axis is indicative of a lack of equivalence. The resulting plots are given in Figure 4.3. It should be noted that only subjects that have a complete set of three values are included in the plots. As in Chapter 3, we also include the centroids and the convex hulls. From this plot there appears to be little separation of the centroids in the vertical direction. It seems likely that T and R are average bioequivalent.

Of course to determine if T and R are sufficiently similar to each other to be declared ABE, we must apply the TOST procedure. The results are given in Tables 4.3, where subjects have been fitted as fixed effects. We can see that T and R are clearly average bioequivalent.

Example 4.2

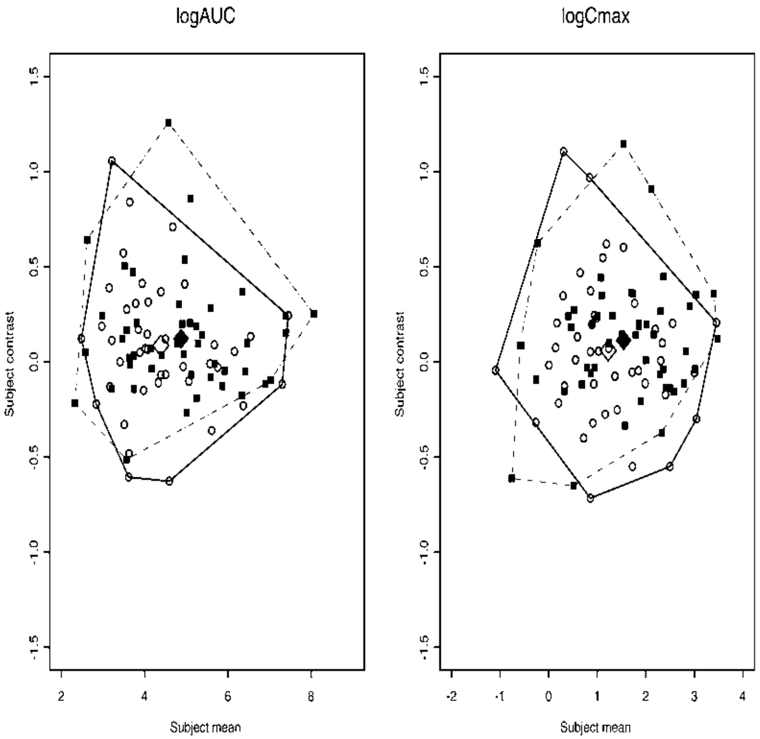

Figure 4.3 *Example 4.1: Subject Contrasts versus Means Plot*

The data in Tables 4.27 and 4.28 are from a trial that also used the sequences RTT and TRR. The corresponding subject profiles are given in Figure 4.4. Relatively large between-subject variation is evident, with perhaps a higher variance on the logAUC scale. It is not clear if, on average, T is giving higher or lower values than R. The group-by-period means are given in Table 4.4, where because of the missing data, we have again indicated the number of subjects that provided data for each mean. These are plotted in Figure 4.5, where the upper line in each plot refers to Sequence RTT. Even allowing for the difference in absolute size of the logAUC and logCmax means, there is a different pattern of formulation differences within each period for the two metrics. There appears to be more of a difference between the formulations on the logAUC scale. The only other notable feature is that the means for Sequence 1 are consistently higher than the corresponding means for Sequence 2.

A better impression of the difference, if any, between T and R is ob-

Table 4.3 *Example 4.1: TOST Procedure Results*

log scale		
Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% Confidence Interval
logAUC	-0.0270	(-0.1395, 0.0855)
logCmax	-0.0557	(-0.1697, 0.0583)

back-transformed		
Endpoint	$\exp(\hat{\mu}_T - \hat{\mu}_R)$	90% Confidence Interval
AUC	0.97	(0.87, 1.09)
Cmax	0.95	(0.84, 1.06)

tained from a plot of the subject contrasts against the subject means. For Example 4.2, this is given as Figure 4.6. There is a clear separation of the convex hulls for both metrics suggesting a lack of bioequivalence. In addition, there is clearly more variability in the plotted points from Sequence 2 as compared to Sequence 1.

The results of applying the TOST procedure to these data are given in Table 4.5. Insufficient evidence were present to conclude that T and R are ABE.

4.3 Within-subject Variability

It is clear that each of our possible designs for three periods has T repeated in one sequence and R repeated in the other. It is therefore possible to separately estimate the within-subject variance of T and the within-subject variance of R. We will denote these by σ^2_{WT} and σ^2_{WR}, respectively. Let us concentrate on the design that has sequences RTT and TRR. Suppose we want an estimate of σ^2_{WT} for the logAUC values. A simple method of estimation uses only the subjects that have a logAUC value on both occurrences of T. Suppose we denote these values by y_{12k} and y_{13k}, for such a subject k in the first sequence group. Then $\text{Var}(y_{12k} - y_{13k}) = 2\sigma^2_{WT}$. To estimate this variance we first construct the set of differences $y_{12k} - y_{13k}$ and then estimate the variance of the differences. The estimate so obtained, and divided by 2, gives $\hat{\sigma}^2_{WT}$. A similar process can be used to calculate $\hat{\sigma}^2_{WR}$ using the appropriate subjects in the second sequence group. In a 2 × 2 cross-over a similar

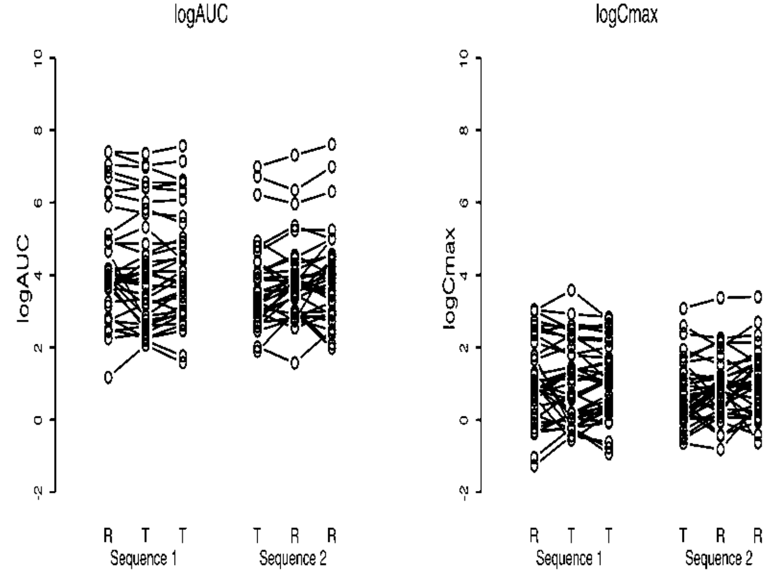

Figure 4.4 *Example 4.2: Subject Profiles Plot*

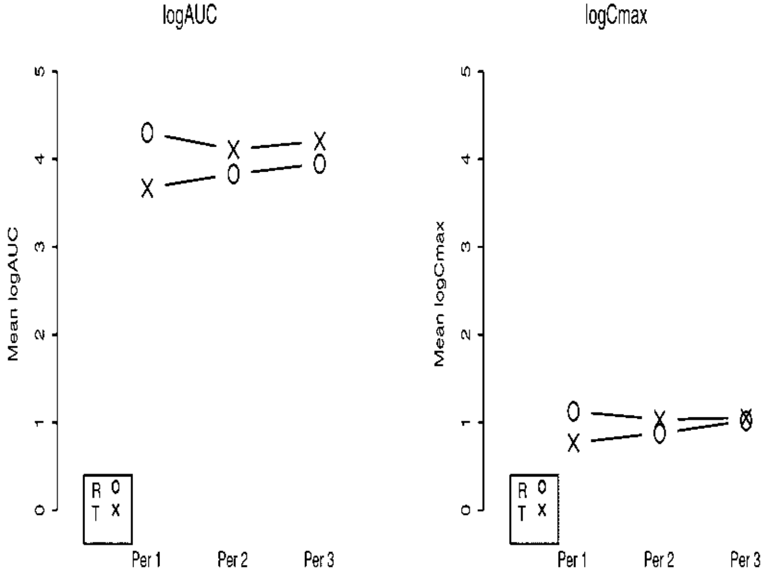

Figure 4.5 *Example 4.2: Groups-by-periods Plot*

Table 4.4 *Example 4.2: Group-by-Period Means (sample size in brackets)*

Group	Period 1	Period 2	Period 3	Mean
		logAUC		
1(RTT)	$\bar{y}_{11.} = 4.30(37)$	$\bar{y}_{12.} = 4.11(38)$	$\bar{y}_{13.} = 4.21(38)$	$\bar{y}_{1..} = 4.21$
2(TRR)	$\bar{y}_{21.} = 3.67(33)$	$\bar{y}_{22.} = 3.83(34)$	$\bar{y}_{23.} = 3.95(35)$	$\bar{y}_{2..} = 3.82$
Mean	$\bar{y}_{.1.} = 4.01$	$\bar{y}_{.2.} = 3.98$	$\bar{y}_{.3.} = 4.08$	$\bar{y}_{...} = 4.02$
		logCmax		
1(RTT)	$\bar{y}_{11.} = 1.13(39)$	$\bar{y}_{12.} = 1.03(39)$	$\bar{y}_{13.} = 1.05(39)$	$\bar{y}_{1..} = 1.07$
2(TRR)	$\bar{y}_{21.} = 0.77(35)$	$\bar{y}_{22.} = 0.88(35)$	$\bar{y}_{23.} = 1.02(35)$	$\bar{y}_{2..} = 0.89$
Mean	$\bar{y}_{.1.} = 0.96$	$\bar{y}_{.2.} = 0.96$	$\bar{y}_{.3.} = 1.04$	$\bar{y}_{...} = 0.98$

Table 4.5 *Example 4.2: TOST Procedure Results*

Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% Confidence Interval
logAUC	-0.1719	(-0.2630, -0.0809)
logCmax	-0.1299	(-0.2271, -0.0327)

Endpoint	$\exp(\hat{\mu}_T - \hat{\mu}_R)$	90% Confidence Interval
AUC	0.84	(0.77, 0.92)
Cmax	0.88	(0.80, 0.97)

procedure is used to calculate $\hat{\sigma}^2_W$ under the assumption that $\sigma^2_{WT} = \sigma^2_{WR} = \sigma^2_W$ [237].

Doing this for Example 4.1, we get, for logAUC: $\hat{\sigma}^2_{WR} = 0.168$ and $\hat{\sigma}^2_{WT} = 0.396$, and for logCmax: $\hat{\sigma}^2_{WR} = 0.214$ and $\hat{\sigma}^2_{WT} = 0.347$.

For Example 4.2, the corresponding values for logAUC are $\hat{\sigma}^2_{WR} = 0.168$, $\hat{\sigma}^2_{WT} = 0.065$, and for logCmax are $\hat{\sigma}^2_{WR} = 0.201$ and $\hat{\sigma}^2_{WT} = 0.087$.

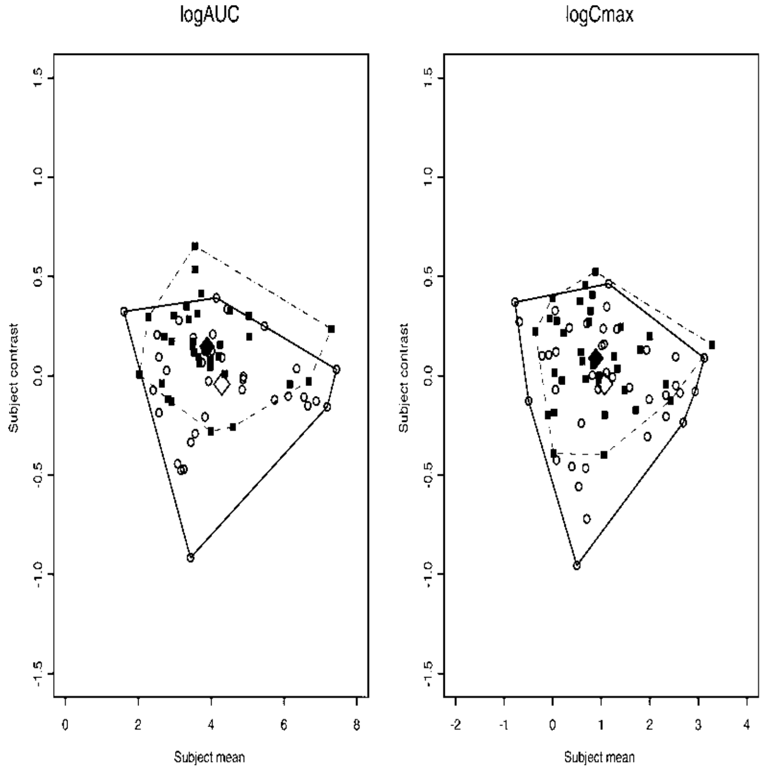

Figure 4.6 *Example 4.2: Subject Contrasts versus Means Plot*

In both examples, $\hat{\sigma}^2_{WR} > 0.09$ for each metric, indicating that the Reference formulations are highly variable.

In Chapter 5 we will give an alternative method of estimation.

4.4 Robust Analyses for Three Period Designs

The model assumed for our data (logAUC or logCmax) is as given in (3.1) in Chapter 3:

$$y_{ijk} = \mu_{d[i,j]} + \pi_j + \gamma_i + \xi_{k(i)} + \varepsilon_{ijk}.$$

This makes some strong assumptions about the variance and covariance structure of the repeated measurements on each subject. In particular, it assumes that the variance of each repeated measurement is the same and that the covariance between any two repeated measurements is the same, i.e.,$\text{Var}(y_{ijk}) = \sigma_B^2 + \sigma_W^2$ and

$$\text{Cov}(y_{i1k}, y_{i2k}) = \text{Cov}(y_{i1k}, y_{i3k}) = \text{Cov}(y_{i2k}, y_{i3k}) = \sigma_B^2.$$

If there is any doubt that these assumptions are unlikely to be true, an alternative, robust, analysis is possible. The analysis is robust in the sense that the only assumptions made are that: the responses from different subjects are independent, the two groups of subjects are a random sample from the same statistical population, and that the period, treatment, and other effects act additively. The analysis, for the sequences RTT and TRR uses the same subject contrasts that were used to construct the subject contrasts versus means plot: $d_{ik} = -(2y_{i1k} - y_{i2k} - y_{i3k})/4$, where it will be recalled that $\hat{\delta} = \bar{d}_{1.} - \bar{d}_{2.}$. The assumptions made in the analysis are then those referring to d_{ik}: the values from different subjects are independent, the values in each group are a random sample from the same statistical population and finally the only difference, if any, between the groups is a shift the value of the mean (or median).

Having calculated the values of d_{ik} (for those patients that provided three repeated measurements), the TOST analysis uses the 90% confidence interval based on the t-distribution or, if the data are very non-normal, the Hodges-Lehmann version of the confidence interval.

For the kth subject in Group 1, $k = 1, 2, \ldots, n_1$, we define $d_{1k} = -(2y_{11k} - y_{12k} - y_{13k})/4$ in Group 1 and $d_{2k} = -(2y_{21k} - y_{22k} - y_{23k})/4$ in Group 2. If ${\sigma_d}^2 = \text{Var}[d_{1k}] = \text{Var}[d_{2k}]$, then

$$\text{Var}[\hat{\delta}] = \sigma_d^2 \left[\frac{1}{n_1} + \frac{1}{n_2} \right].$$

To estimate ${\sigma_d}^2$ we use the usual pooled estimator

$${s_p}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)},$$

where s_1^2 is the sample variance of d_{1k} and ${s_2}^2$ is the sample variance of d_{2k}. To construct the 90% confidence interval for δ we use that fact that when $\delta = 0$,

$$t = \frac{\bar{d}_{1.} - \bar{d}_{2.}}{\left[s_p^2 (\frac{1}{n_1} + \frac{1}{n_2}) \right]^{\frac{1}{2}}}$$

has the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. It will be noted that the degrees of freedom for the usual TOST interval, based on subjects with all three repeated measurements is $2(n_1 + n_2) - 3$, as compared to $(n_1 + n_2 - 2)$ for the robust method. Even so, this loss of degrees of freedom rarely has a major effect on the conclusions.

For the data in Example 4.1, $n_1 = 42$, $n_2 = 46$ for logAUC, and $\bar{d}_{1.} = -0.0850$, $\bar{d}_{2.} = -0.1215$ and $\hat{\delta} = -0.0365$. Further, $s_1^2 = 0.1191$, $s_2^2 = 0.0917$, and $s_p^2 = 0.1048$. Based on the t-distribution with 86 degrees of freedom, the TOST 90% confidence is (-0.1514, 0.0783). If the usual analysis (of y_{ijk}) is done, the corresponding interval is (-0.1516, 0.0785), based on the t-distribution with 173 degrees of freedom. The robust interval is a little wider but, in this case at least, the conclusions are the same. For logCmax the robust confidence interval is (-0.1712, 0.0651), on 91 degrees of freedom, as compared to the usual interval of (-0.1685, 0.0623) on 183 degrees of freedom.

For the data in Example 4.2, $n_1 = 35$, $n_2 = 32$ for logAUC and the robust interval is $(-0.2880, -0.0897)$ on 65 df and the usual interval is $(-0.2800, -0.0977)$ on 131 df. For logCmax, $n_1 = 39$, $n_2 = 35$ and the robust interval is (-0.2359, -0.0239) on 72 df and the usual interval is $(-0.2271, -0.0327)$ on 145 df. Again, the conclusions from both approaches are the same.

An alternative confidence interval that does not rely on the t-distribution is the Hodges-Lehmann point confidence interval described in Chapter 3. In the notation of that chapter, we let $X_k = d_{2k}$ and $Y_k = d_{1k}$. The resulting confidence intervals for Example 4.1 are $(-0.1196, -0.0731)$ for logAUC and $(-0.1679, 0.0394)$ for logCmax. For Example 4.2, the corresponding intervals are $(-0.2635, -0.0771)$ for logAUC and $(-0.2042, -0.0071)$ for logCmax. The conclusions obtained above are not changed for either example. The Hodges-Lehmann confidence intervals can also be constructed using StatXact. For Example 4.1 these are (-0.1199, 0.0734) for logAUC and (-0.1683, 0.0410) for logCmax. For Example 4.2 these are (-0.2635, -0.0771) and (-0.2049, -0.0070), respectively.

4.5 Four-period Designs

4.5.1 Choice of Design

As already mentioned, four-period designs are recommended by the FDA when the reference drug is highly variable (i.e., $\sigma_W^2 > 0.09$). If we discard the sequences RRRR and TTTT then there are seven different two-sequence designs and they are given below:

1.	R	R	T	T	2.	R	T	R	T	3.	R	T	T	R
	T	T	R	R		T	R	T	R		T	R	R	T
4.	R	T	R	R	5.	R	R	T	R	6.	R	T	T	T
	T	R	T	T		T	T	R	T		T	R	R	R

and

7.	R	R	R	T
	T	T	T	R

The efficiencies of these designs are given in Table 4.6. In the presence of unequal carry-over effects, only Designs 1 and 3 are worth consideration [237]. It is worthy of note that the design recommended by the FDA is Design 2. In the absence of a difference in carry-over effects Design 1, 2 and 3 are equally, and fully, efficient.

Table 4.6 *Efficiencies of Designs 1 through 7*

Design	Adjusted for Carry-over	Unadjusted for Carry-over
1	90.91	100.00
2	18.18	100.00
3	90.91	100.00
4	54.55	75.00
5	54.55	75.00
6	72.73	75.00
7	66.67	75.00

4.5.2 Examples of Data Analysis for Four-period Designs

Example 4.3
The data in Table 4.29 are from a trial with four periods that used the sequences RTTR and TRRT. This trial was quite small with 8 subjects in Group 1 and 9 in Group 2. The subject profiles plots for logAUC and logCmax are given in Figure 4.7. From this plot it is difficult to discern if T and R are ABE. The group-by-period means are given in Table 4.7 and are plotted in Figure 4.8. These seem to indicate that T and R are ABE.

The subject contrasts plots are given in Figure 4.9 and reveal a difference in the centroids, particularly for logCmax, although the actual size of the difference is relatively small. There is also some evidence that there is more variability in the logAUC contrasts for the subjects on sequence TRRT. To clarify matters regarding ABE we refer to the results of the TOSTs given in Table 4.8, where the fixed-subjects models have been fitted. The evidence is in favor of concluding that T and R are ABE, although for logCmax the lower end of the confidence interval

Table 4.7 *Example 4.3: Group-by-Period Means (sample size in brackets)*

		logAUC			
Group	Period 1	Period 2	Period 3	Period 4	Mean
1(RTTR)	$\bar{y}_{11.} = 8.94(8)$	$\bar{y}_{12.} = 8.99(8)$	$\bar{y}_{13.} = 8.96(8)$	$\bar{y}_{14.} = 8.91(8)$	$\bar{y}_{1..} = 8.95$
2(TRRT)	$\bar{y}_{21.} = 8.83(9)$	$\bar{y}_{22.} = 8.80(9)$	$\bar{y}_{23.} = 8.85(9)$	$\bar{y}_{24.} = 8.89(8)$	$\bar{y}_{2..} = 8.84$
Mean	$\bar{y}_{.1.} = 8.88$	$\bar{y}_{.2.} = 8.89$	$\bar{y}_{.3.} = 8.91$	$\bar{y}_{.4.} = 8.90$	$\bar{y}_{...} = 8.89$
		logCmax			
1(RTTR)	$\bar{y}_{11.} = 7.02(8)$	$\bar{y}_{12.} = 7.00(8)$	$\bar{y}_{13.} = 6.98(8)$	$\bar{y}_{14.} = 7.08(8)$	$\bar{y}_{1..} = 7.02$
2(TRRT)	$\bar{y}_{21.} = 6.83(9)$	$\bar{y}_{22.} = 7.02(9)$	$\bar{y}_{23.} = 7.06(9)$	$\bar{y}_{24.} = 7.02(8)$	$\bar{y}_{2..} = 6.98$
Mean	$\bar{y}_{.1.} = 6.92$	$\bar{y}_{.2.} = 7.01$	$\bar{y}_{.3.} = 7.02$	$\bar{y}_{.4.} = 7.05$	$\bar{y}_{...} = 7.00$

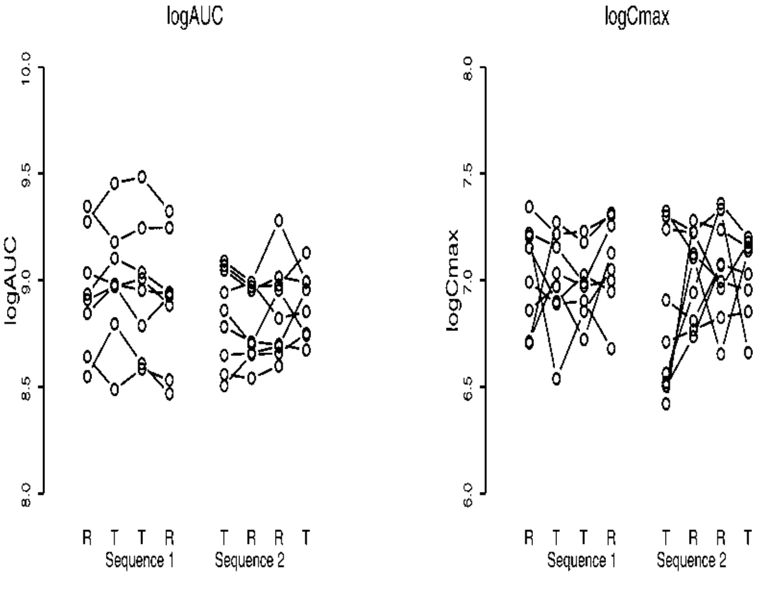

Figure 4.7 *Example 4.3: Subject Profiles Plot*

is close to the lower boundary of -0.2231 (on the log scale). The robust and Hodges-Lehmann exact confidence intervals are (-0.0080, 0.0811) and $(-0.0148, 0.0834)$, respectively, for logAUC and are (-0.1786, 0.0038) and (-0.2001, 0.0073), respectively, for logCmax.

Example 4.4

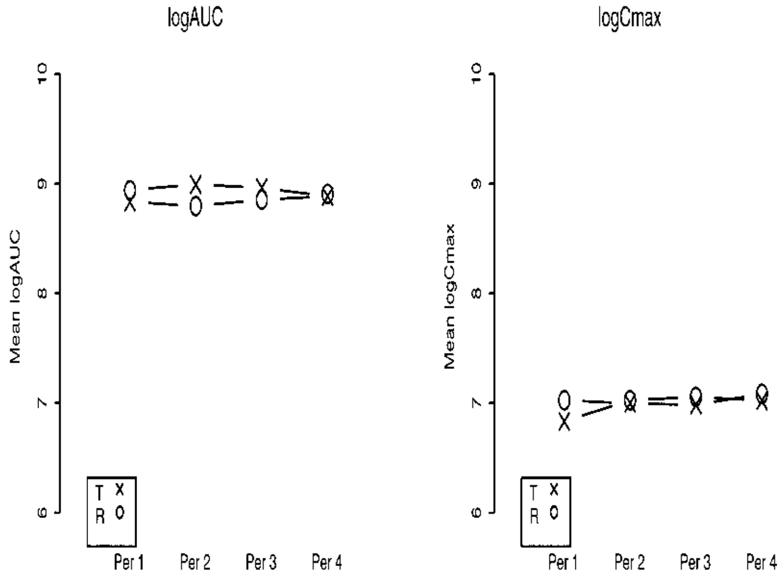

Figure 4.8 *Example 4.3: Groups-by-periods Plot*

Table 4.8 *Example 4.3: TOST Procedure Results*

Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% Confidence Interval
logAUC	0.0352	(-0.0044, 0.0748)
logCmax	-0.0963	(-0.1881, 0.0045)
Endpoint	$\exp(\hat{\mu}_T - \hat{\mu}_R)$	90% Confidence Interval
AUC	1.04	(1.00, 1.08)
Cmax	0.91	(0.83, 1.00)

The data in Tables 4.30 and 4.31 are from another four-period design, but this time the sequences used were RTRT and TRTR. The subject profiles plots are given in Figure 4.10. The large number of subjects per group makes it difficult to discern much from this plot, other than the relatively large between-subject variation. The group-by-period means are given in Table 4.9 and plotted in Figure 4.11. The picture is clearer now, with a suggestion that for logCmax that T and R might not be

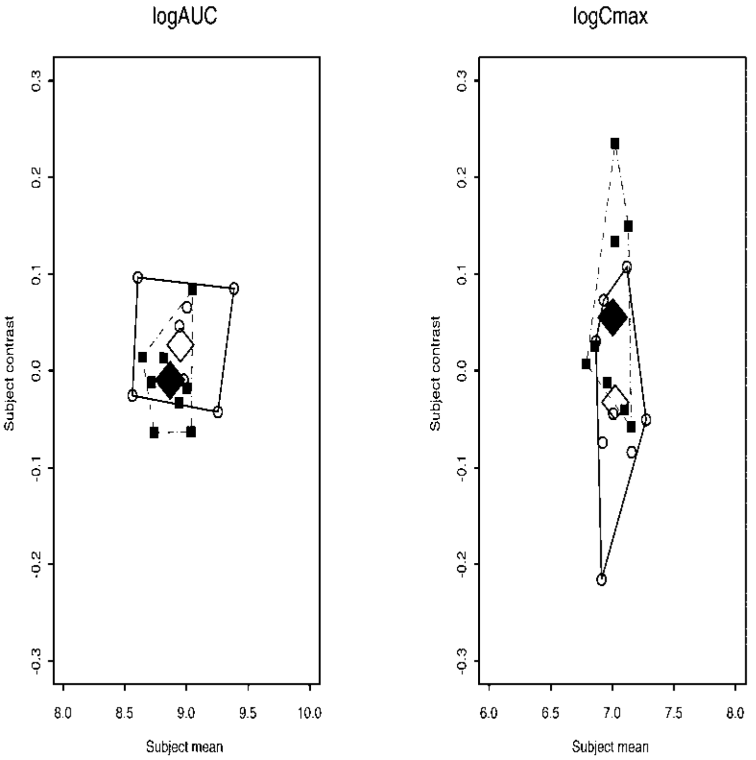

Figure 4.9 *Example 4.3: Subject Contrasts versus Means Plot*

ABE. The subject contrasts versus means plot is given in Figure 4.12, where it is clear there is a relatively large vertical gap in the centroids for logCmax. This is confirmed from the TOST results given in Table 4.10, here the lower bound of the 90% confidence interval for logCmax is a long way above 0.2231, upper the regulatory limit. The robust and Hodges-Lehmann confidence intervals for logAUC are (0.0311, 1758) and (0.0256, 0.1630), respectively. The corresponding intervals for logCmax are (0.2685, 0.5623) and (0.2681, 0.5626). There is very strong evidence that T and R are not ABE.

4.6 Designs with More Than Two Treatments

As already mentioned in the Introduction, designs for more than two treatments may be used to show bioequivalence, but these are less com-

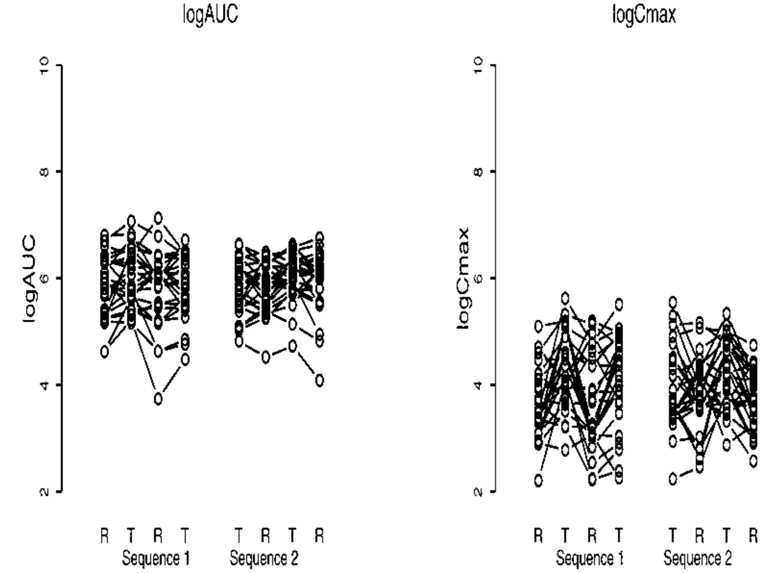

Figure 4.10 *Example 4.4: Subject Profiles Plot*

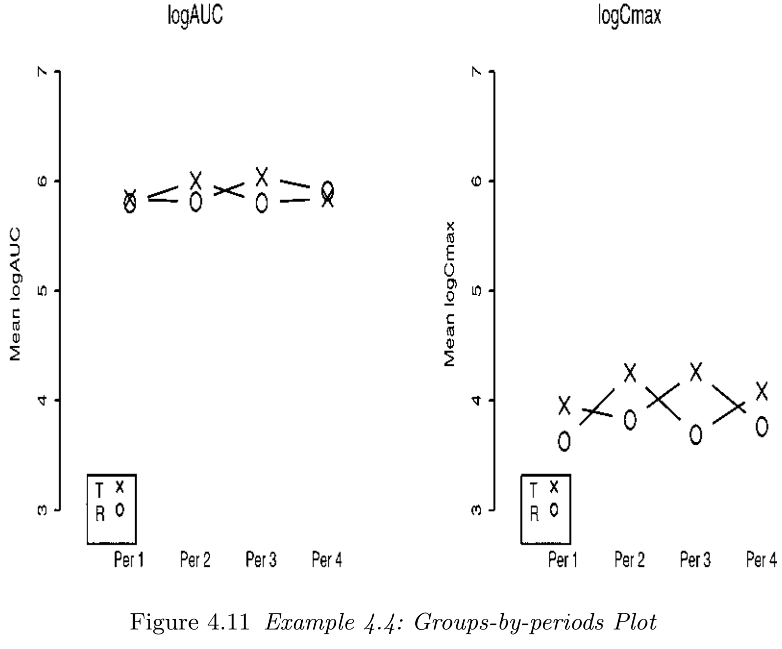

Figure 4.11 *Example 4.4: Groups-by-periods Plot*

Table 4.9 *Example 4.4: Group-by-Period Means (sample size in brackets)*

		logAUC			
Group	Period 1	2	3	4	Mean
1	$\bar{y}_{11.} = 5.80(27)$	$\bar{y}_{12.} = 6.00(27)$	$\bar{y}_{13.} = 5.80(26)$	$\bar{y}_{14.} = 5.85(26)$	$\bar{y}_{1..} = 5.86$
2	$\bar{y}_{21.} = 5.84(27)$	$\bar{y}_{22.} = 5.81(27)$	$\bar{y}_{23.} = 6.04(26)$	$\bar{y}_{24.} = 5.91(26)$	$\bar{y}_{2..} = 5.90$
Mean	$\bar{y}_{.1.} = 5.82$	$\bar{y}_{.2.} = 5.91$	$\bar{y}_{.3.} = 5.92$	$\bar{y}_{.4.} = 5.88$	$\bar{y}_{...} = 5.88$
		logCmax			
1	$\bar{y}_{11.} = 3.63(27)$	$\bar{y}_{12.} = 4.26(27)$	$\bar{y}_{13.} = 3.69(26)$	$\bar{y}_{14.} = 4.09(26)$	$\bar{y}_{1..} = 3.91$
2	$\bar{y}_{21.} = 3.96(27)$	$\bar{y}_{22.} = 3.82(27)$	$\bar{y}_{23.} = 4.26(26)$	$\bar{y}_{24.} = 3.76(26)$	$\bar{y}_{2..} = 3.95$
Mean	$\bar{y}_{.1.} = 3.79$	$\bar{y}_{.2.} = 4.04$	$\bar{y}_{.3.} = 3.97$	$\bar{y}_{.4.} = 3.93$	$\bar{y}_{...} = 3.93$
		Group 1=RTRT; 2=TRTR			

Table 4.10 *Example 4.4: TOST Procedure Results*

Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% Confidence Interval
logAUC	0.1002	(0.0289, 0.1715)
logCmax	0.4140	(0.2890, 0.5389)
Endpoint	$\exp(\hat{\mu}_T - \hat{\mu}_R)$	90% Confidence Interval
AUC	1.11	(1.03, 1.19)
Cmax	1.51	(1.34, 1.71)

mon than those for two treatments. Examples 4.5 and 4.6, below are examples where three and four treatments, respectively, were used.

Example 4.5. Trial with Three Treatments.
In this trial there were two 'reference' formulations, R and S, where R was a dose made up of three 100 mg tablets and S was a dose made up of a 200 mg tablet and a 100 mg tablet. The test formulation was a single 300 mg tablet. Two reference formulations were used because the 200 mg tablet was not available in the early stages of the confirmatory trial when the 3×100 mg dose was used. The aim of the trial was to show that T and R were ABE **and** T and S were ABE. The subjects in the trial were randomly allocated to the six sequences: RST, RTS, SRT, STR,

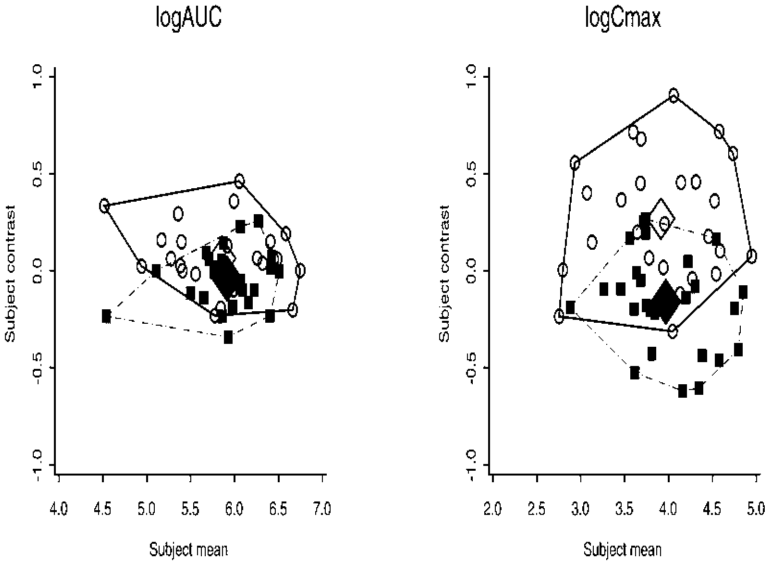

Figure 4.12 *Example 4.4: Subject Contrasts versus Means Plot*

TRS and TSR. The data from this trial are given in Tables 4.32, 4.33, and 4.34. This design is known as a Williams design (see [237], Chapter 4) and is balanced for carry-over effects. In the presence of carry-over effects the variance of any pairwise difference between the formulations is $(5\sigma_W^2)/(12r)$, where r is the number of replications of the complete set of six sequences. In the absence of carry-over effects this variance is $(\sigma_W^2)/(3r)$, which is also the variance in an ideal design. Hence, the efficiency of the Williams design for three formulations is 80% in the presence of carry-over effects and is 100% in the absence of carry-over effects. Of course, we do not expect to see any differential carry-over effects and, as we shall see, there is no suggestion from the data that such effects need concern us.

The subject profiles plots are given in Figures 4.13, 4.14, and 4.15. Large between-subject variation is evident and there is a suggestion that S gives a higher response than R or T. The group-by-period means are given in Table 4.11 and are plotted in Figure 4.16. There is a clear ordering within all but one of the periods with S giving the highest mean response and R the lowest. To determine if each of T and R and T and S are ABE, we use the TOST procedure for each difference and the results are given in Table 4.12. Note that we do not adjust for multiple testing as we require both pairs to be ABE. We can conclude that T is

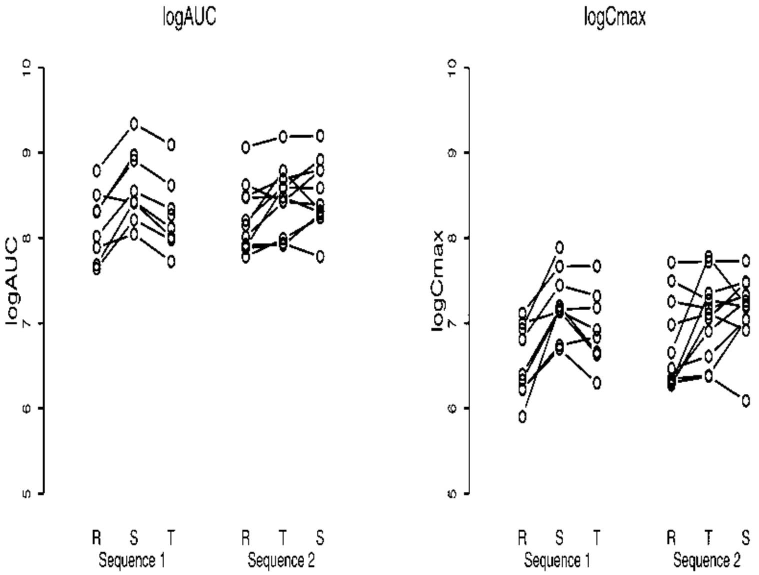

Figure 4.13 *Example 4.5: Subject Profiles Plot: Sequences 1 and 2*

not ABE to R and S. The ordering of the formulation means is, as noted from the previous plots, that S gives a significantly higher response than T, which in turn is significantly higher than R.

Example 4.6. Trial with Four Treatments

In this trial the test formulation could be given as a low or a high dose. Hence, it was necessary to compare these with low and high doses, respectively, of the reference formulation. The four formulations were labelled as A, B, C, and D, where A is the Reference, Low dose, B is the Test, Low dose, C is the Reference, High dose, and D is the Test, High dose. The comparisons of interest were therefore B-A and D-C. A Williams design for four periods was used in the trial with sequences ADBC, BADC, CBDA and DCAB. The efficiency of this design is 90.91% in the presence of differential carry-over effects and is 100% in their absence.

The data from this trial are given in Tables 4.35 to 4.38 and the subject profiles plots are given in Figures 4.17 and 4.18. The large changes in the plots occur when moving from a low to a high dose and vice versa. Within a dose there seems relatively good agreement between T and R. The group-by-period means are given in Table 4.13 and are plotted in Figure 4.19, where it will be noted that the symbols for A and C are the circle and triangle, respectively, and the symbols for B and D are the vertical and diagonal crosses, respectively. The large difference between

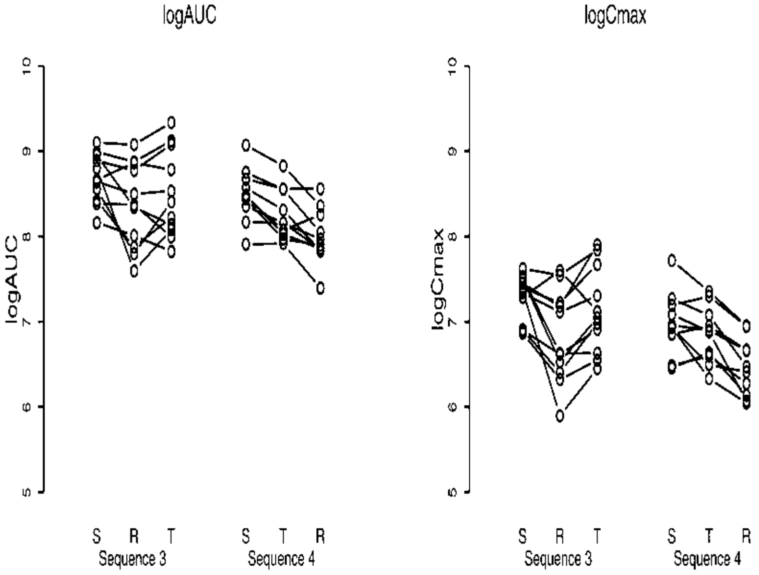

Figure 4.14 *Example 4.5: Subject Profiles Plot: Sequences 3 and 4*

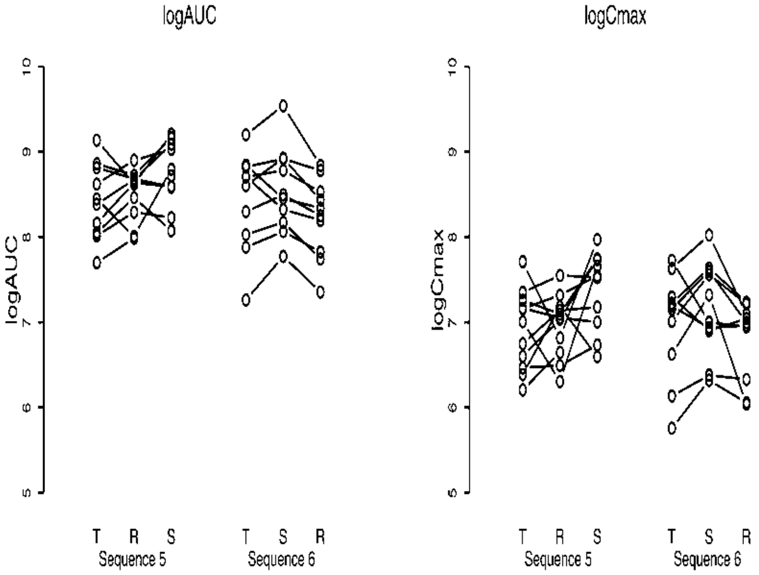

Figure 4.15 *Example 4.5: Subject Profiles Plot: Sequences 5 and 6*

Table 4.11 *Example 4.5: Group-by-Period Means (sample size in brackets)*

		logAUC		
Group	Period 1	Period 2	Period 3	Mean
1(RST)	$\bar{y}_{11.} = 8.14(8)$	$\bar{y}_{12.} = 8.61(8)$	$\bar{y}_{13.} = 8.27(8)$	$\bar{y}_{1..} = 8.34$
2(RTS)	$\bar{y}_{21.} = 8.23(11)$	$\bar{y}_{22.} = 8.45(11)$	$\bar{y}_{23.} = 8.51(11)$	$\bar{y}_{2..} = 8.40$
3(SRT)	$\bar{y}_{31.} = 8.69(11)$	$\bar{y}_{32.} = 8.37(11)$	$\bar{y}_{33.} = 8.50(11)$	$\bar{y}_{3..} = 8.52$
4(STR)	$\bar{y}_{41.} = 8.49(10)$	$\bar{y}_{42.} = 8.25(10)$	$\bar{y}_{43.} = 8.00(10)$	$\bar{y}_{4..} = 8.25$
5(TRS)	$\bar{y}_{51.} = 8.42(10)$	$\bar{y}_{52.} = 8.50(10)$	$\bar{y}_{53.} = 8.75(10)$	$\bar{y}_{5..} = 8.55$
6(TSR)	$\bar{y}_{61.} = 8.43(10)$	$\bar{y}_{62.} = 8.54(10)$	$\bar{y}_{63.} = 8.23(10)$	$\bar{y}_{6..} = 8.40$
Mean	$\bar{y}_{.1.} = 8.41$	$\bar{y}_{.2.} = 8.45$	$\bar{y}_{.3.} = 8.38$	$\bar{y}_{...} = 8.41$
		logCmax		
1(RST)	$\bar{y}_{11.} = 6.55(9)$	$\bar{y}_{12.} = 7.23(9)$	$\bar{y}_{13.} = 6.94(8)$	$\bar{y}_{1..} = 6.91$
2(RTS)	$\bar{y}_{21.} = 6.74(11)$	$\bar{y}_{22.} = 7.07(11)$	$\bar{y}_{23.} = 7.16(11)$	$\bar{y}_{2..} = 6.99$
3(SRT)	$\bar{y}_{31.} = 7.29(11)$	$\bar{y}_{32.} = 6.82(11)$	$\bar{y}_{33.} = 7.13(11)$	$\bar{y}_{3..} = 7.08$
4(STR)	$\bar{y}_{41.} = 6.99(10)$	$\bar{y}_{42.} = 6.85(10)$	$\bar{y}_{43.} = 6.46(10)$	$\bar{y}_{4..} = 6.77$
5(TRS)	$\bar{y}_{51.} = 6.91(11)$	$\bar{y}_{52.} = 6.97(11)$	$\bar{y}_{53.} = 7.37(10)$	$\bar{y}_{5..} = 7.07$
6(TSR)	$\bar{y}_{61.} = 6.97(10)$	$\bar{y}_{62.} = 7.17(10)$	$\bar{y}_{63.} = 6.79(10)$	$\bar{y}_{6..} = 6.98$
Mean	$\bar{y}_{.1.} = 6.92$	$\bar{y}_{.2.} = 7.01$	$\bar{y}_{.3.} = 6.98$	$\bar{y}_{...} = 6.97$

the means values for the two doses (circle and triangle versus vertical and diagonal cross) is clearly displayed, as is the relatively small difference between T and R within doses (circle versus triangle and vertical versus diagonal cross). At first sight, at least, it appears the T and R are ABE at each dose. The results of the TOST procedure are given in Table 4.14, and these confirm the conclusions made from the plots.

4.7 Nonparametric Analyses of Tmax

There are a number of alternative approaches to developing a distribution-free or nonparametric analysis of data from cross-over trials with three or more treatments. The simplest and most familiar is an extension of the nonparametric analysis of the design with two treatments and sequences: TR/RT. This can only be applied to a subset of designs and is based on a stratified analysis for two treatments, resulting in the Van Elteren test (see [388] for example). The particular designs to which this approach is

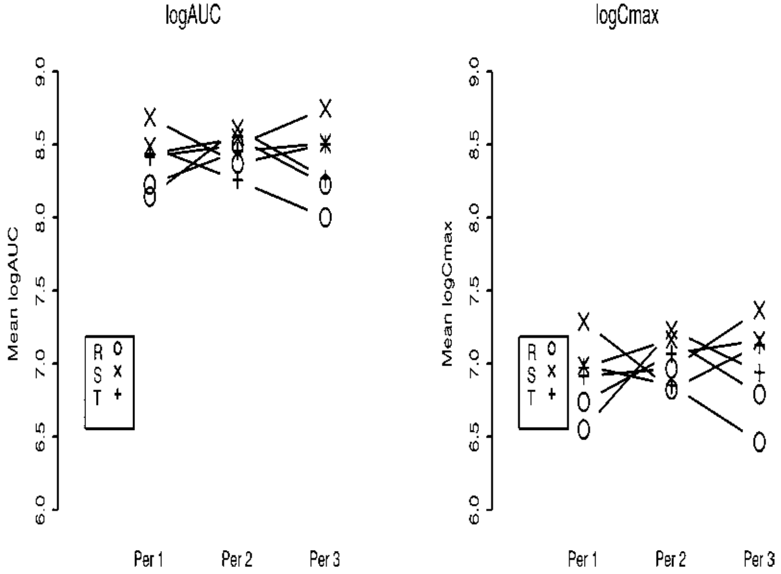

Figure 4.16 *Example 4.5: Groups-by-periods Plot*

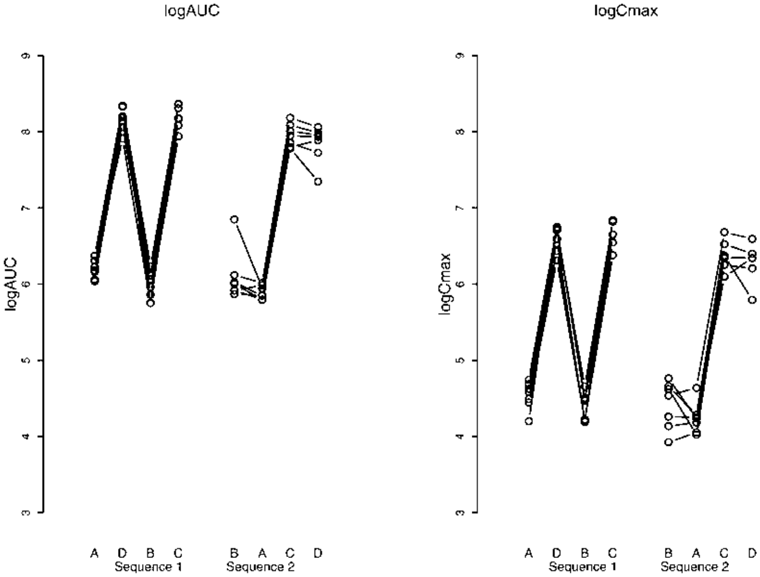

Figure 4.17 *Example 4.6: Subject Profiles Plot: Sequences 1 and 2*

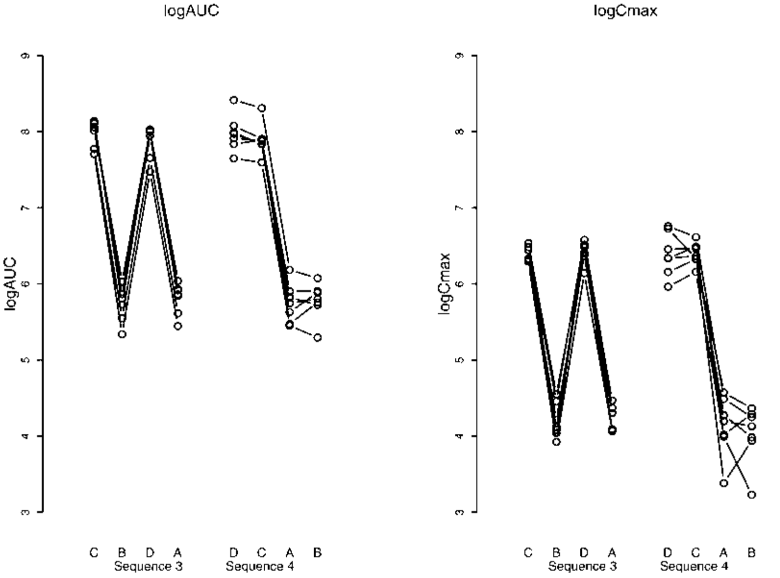

Figure 4.18 *Example 4.6: Subject Profiles Plot: Sequences 3 and 4*

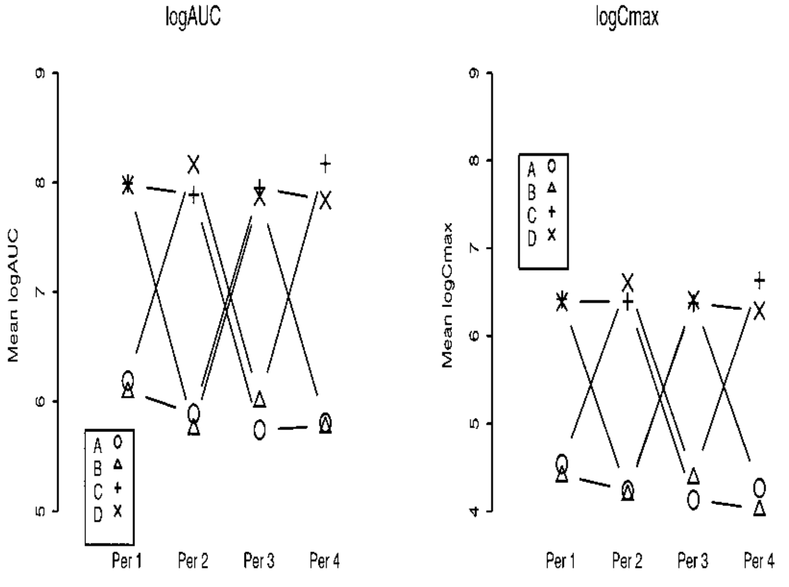

Figure 4.19 *Example 4.6: Groups-by-periods Plot*

Table 4.12 *Example 4.5: TOST Procedure Results*

Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	T-R 90% Confidence Interval
logAUC	0.1505	(0.0865, 0.2145)
logCmax	0.2618	(0.1747, 0.3489)

Endpoint	$\hat{\mu}_T - \hat{\mu}_S$	T-S 90% Confidence Interval
logAUC	-0.1888	(-0.2532, -0.1243)
logCmax	-0.2044	(-0.2921, -0.1167)

Endpoint	$\exp(\hat{\mu}_T - \hat{\mu}_R)$	T-R 90% Confidence Interval
AUC	1.16	(1.09, 1.24)
Cmax	1.30	(1.19, 1.42)

Endpoint	$\exp(\hat{\mu}_T - \hat{\mu}_S)$	T-S 90% Confidence Interval
AUC	0.83	(0.78, 0.88)
Cmax	0.82	(0.75, 0.89)

applicable are those that have embedded within them a suitable set of RT/TR designs. We will illustrate such sets for three treatments in the following subsections. For arbitrary designs, confidence intervals can be derived using bootstrap sampling.

The most common need for a nonparametric analysis of bioequivalence data is in the analysis of Tmax. In the following subsections we will analyse Tmax data collected in the trials described in Examples 4.5 and 4.6.

4.7.1 Three Treatments

The data in Tables 4.15, 4.16, and 4.17 are the Tmax values collected in the trial described in Example 4.5. The design is displayed again in Table 4.18. It can be seen that the six sequences have been arranged into three

Table 4.13 *Example 4.6: Group-by-Period Means (sample size in brackets)*

		logAUC			
Group	Period 1	Period 2	Period 3	Period 4	Mean
1(ADBC)	$\bar{y}_{11.} = 6.19(7)$	$\bar{y}_{12.} = 8.17(7)$	$\bar{y}_{13.} = 6.01(7)$	$\bar{y}_{14.} = 8.17(7)$	$\bar{y}_{1..} = 7.14$
2(BACD)	$\bar{y}_{21.} = 6.10(7)$	$\bar{y}_{22.} = 5.89(7)$	$\bar{y}_{23.} = 7.95(7)$	$\bar{y}_{24.} = 7.84(7)$	$\bar{y}_{2..} = 6.95$
1(CBDA)	$\bar{y}_{11.} = 7.99(7)$	$\bar{y}_{12.} = 5.76(7)$	$\bar{y}_{13.} = 7.87(7)$	$\bar{y}_{14.} = 5.81(7)$	$\bar{y}_{1..} = 6.86$
2(DCAB)	$\bar{y}_{21.} = 7.98(7)$	$\bar{y}_{22.} = 7.89(7)$	$\bar{y}_{23.} = 5.74(7)$	$\bar{y}_{24.} = 5.78(7)$	$\bar{y}_{2..} = 6.85$
Mean	$\bar{y}_{.1.} = 7.06$	$\bar{y}_{.2.} = 6.93$	$\bar{y}_{.3.} = 6.90$	$\bar{y}_{.4.} = 6.90$	$\bar{y}_{...} = 6.95$
		logCmax			
1(ADBC)	$\bar{y}_{11.} = 4.54(7)$	$\bar{y}_{12.} = 6.61(7)$	$\bar{y}_{13.} = 4.39(7)$	$\bar{y}_{14.} = 6.64(7)$	$\bar{y}_{1..} = 5.54$
2(BACD)	$\bar{y}_{21.} = 4.41(7)$	$\bar{y}_{22.} = 4.24(7)$	$\bar{y}_{23.} = 6.37(7)$	$\bar{y}_{24.} = 6.29(7)$	$\bar{y}_{2..} = 5.33$
1(CBDA)	$\bar{y}_{11.} = 6.42(7)$	$\bar{y}_{12.} = 4.20(7)$	$\bar{y}_{13.} = 6.41(7)$	$\bar{y}_{14.} = 4.26(7)$	$\bar{y}_{1..} = 5.32$
2(DCAB)	$\bar{y}_{21.} = 6.39(7)$	$\bar{y}_{22.} = 6.39(7)$	$\bar{y}_{23.} = 4.13(7)$	$\bar{y}_{24.} = 4.03(7)$	$\bar{y}_{2..} = 5.23$
Mean	$\bar{y}_{.1.} = 5.44$	$\bar{y}_{.2.} = 5.36$	$\bar{y}_{.3.} = 5.33$	$\bar{y}_{.4.} = 5.30$	$\bar{y}_{...} = 5.36$

strata. Stratum I includes the two sequences that contain the TR/RT design in Periods 1 and 2, stratum II includes the two sequences that contain the TR/RT design in Periods 1 and 3, and finally stratum III includes the two sequences that contain the TR/RT design in Periods 2 and 3. Within each stratum, T and R can be compared using the Wilcoxon rank sum test, as described in Section 3.8 of Chapter 3. In particular, the Wilcoxon rank sum and its variance for each stratum can be calculated. An overall test of T versus R can then be obtained by taking a weighted average of the three rank sums and dividing it by the square root of an estimate of the variances of the weighted average to produce a test statistic. This will be illustrated shortly. A defining characteristic of the parent design is that the pair of sequences in each stratum have T and R in matching periods: 1 and 2 in stratum I, 1 and 3 in stratum II and 2 and 3 in stratum III. This is so that the period effect can be eliminated from the treatment comparison.

To compare T and S a different arrangement of the design will be needed, as shown in Table 4.19. It can be seen that the sequences in each stratum are a different selection to those used when comparing T and R. At once we can see some disadvantages of this approach: a design containing the appropriate stratification must be available and a new arrangement of sequences is needed for each individual treatment comparison. A general approach applicable to an arbitrary design will be described later.

Table 4.14 *Example 4.6: TOST Procedure Results*

	B-A	
Endpoint	$\hat{\mu}_B - \hat{\mu}_A$	90% Confidence Interval
logAUC	0.0047	(-0.0544, 0.0638)
logCmax	-0.0355	(-0.1171, 0.0461)

	D-C	
Endpoint	$\hat{\mu}_D - \hat{\mu}_C$	90% Confidence Interval
logAUC	-0.0362	(-0.0953, 0.0230)
logCmax	-0.0301	(-0.1117, 0.0515)

	B-A	
Endpoint	$\exp(\hat{\mu}_B - \hat{\mu}_A)$	90% Confidence Interval
AUC	1.00	(0.95, 1.07)
Cmax	0.97	(0.89, 1.05)

	D-C	
Endpoint	$\exp(\hat{\mu}_D - \hat{\mu}_C)$	90% Confidence Interval
AUC	0.96	(0.91, 1.02)
Cmax	0.97	(0.89, 1.05)

Table 4.15: Example 4.5: Tmax, Williams Design for 3 Treatments

	Sequence TRS				Sequence RTS		
	Period				Period		
Subject	1	2	3	Subject	1	2	3
6	4.00	4.00	2.65	2	3.00	3.00	3.00
12	4.00	4.02	3.02	11	2.97	2.00	2.98
17	2.98	3.98	3.98	16	4.00	3.00	3.00
19	3.98	1.50	2.50	20	3.00	2.02	2.50
29	3.02	3.98	4.00	27	2.00	3.98	2.50
32	2.00	1.98	3.00	31	2.48	1.50	1.48
42	3.00	4.00	2.02	40	1.97	1.50	1.53
46	3.00	3.98	2.98	43	4.02	3.98	3.03
R=3 × 100mg, S=200mg + 100mg, T=Test							

Table 4.15: Example 4.5: Tmax, Williams Design for 3 Treatments

	Sequence TRS				Sequence RTS		
	Period				Period		
Subject	1	2	3	Subject	1	2	3
49	1.98	2.50	1.52	53	2.50	3.98	3.00
60	1.50	3.98	3.00	59	3.00	3.00	3.98
				61	4.00	2.00	4.00
R=3 × 100mg, S=200mg + 100mg, T=Test							

Table 4.16: Example 4.5: Tmax, Williams Design for 3 Treatments

	Sequence TSR				Sequence RST		
	Period				Period		
Subject	1	2	3	Subject	1	2	3
4	2.50	2.98	3.02	9	2.98	2.50	2.50
7	2.48	2.50	3.97	13	2.00	2.98	1.50
14	2.98	3.00	3.00	21	2.52	2.50	1.55
23	1.00	2.98	3.00	28	2.50	2.98	2.97
26	4.05	2.98	6.00	33	2.97	1.52	1.02
36	2.98	3.98	3.00	44	4.00	4.00	3.97
39	4.08	4.00	3.98	50	3.98	4.00	4.00
48	1.03	2.00	2.02	58	3.00	4.00	2.48
54	2.48	2.50	2.50				
56	1.50	1.98	2.48				
R=3 × 100mg, S=200mg + 100mg, T=Test							

Table 4.17: Example 4.5: Tmax, Williams Design for 3 Treatments

	Sequence STR				Sequence SRT		
	Period				Period		
Subject	1	2	3	Subject	1	2	3
5	2.50	1.98	2.55	1	2.50	4.02	3.00
10	1.48	1.50	2.50	8	1.98	1.98	4.00
18	3.00	2.50	2.50	15	1.48	2.50	3.98
22	4.02	3.02	4.02	24	3.00	4.00	4.02
30	4.10	3.02	3.98	25	2.48	3.00	2.98
34	4.12	4.00	3.98	35	2.97	3.98	2.50
37	2.98	1.48	4.02	41	3.03	3.05	3.98
47	2.50	3.00	4.00	45	1.53	4.03	3.03
R=3 × 100mg, S=200mg + 100mg, T=Test							

Table 4.17: Example 4.5: Tmax, Williams Design for 3 Treatments

	Sequence STR				Sequence SRT		
	Period				Period		
Subject	1	2	3	Subject	1	2	3
52	3.00	4.00	2.52	51	3.02	6.00	2.52
55	3.00	3.98	2.48	57	3.00	3.98	3.00
				62	2.98	4.00	2.50
R=3 × 100mg, S=200mg + 100mg, T=Test							

Bioequivalence testing is based on the 90% confidence for the Test versus Reference comparison (on the log scale). However, to motivate and explain the construction of the confidence interval we first start with the construction of the statistic for testing the null hypothesis that the mean treatment difference is zero. We will do this first for a single stratum and then give the generalization.

Table 4.18 *Williams Design for Three Treatments: Stratified for Comparing T and R*

Stratum	Group	Period 1	Period 2	Period 3
I	1	**T**	**R**	S
I	2	**R**	**T**	S
II	3	**T**	S	**R**
II	4	**R**	S	**T**
III	5	S	**T**	**R**
III	6	S	**R**	**T**

Single Stratum

The test statistic is Q, as used by Tudor and Koch [434] for stratified samples and where the variate is the within-stratum ranks of the responses. We first define Q for a single stratum and show its equivalence to the Wilcoxon rank-sum test. In the process we will also show how the Wilcoxon rank-sum can be expressed in terms of U-statistics; this will be useful when we consider the calculation a confidence interval for the difference of T and R.

For a single stratum we assume that there are two sequences TR and RT with T and R in corresponding periods in the two sequences. In

Table 4.19 *Williams Design for Three Treatments: Stratified for Comparing T and S*

Stratum	Group	Period 1	Period 2	Period 3
I	1	**T**	R	**S**
I	6	**S**	R	**T**
II	5	**S**	**T**	R
II	3	**T**	**S**	R
III	2	R	**T**	**S**
III	4	R	**S**	**T**

addition, we assume that the period 1 - period 2 differences have been calculated and ranked (over the total set of differences). Then:

$$Q = \frac{[\bar{R}_1 - \bar{R}_2]^2}{\hat{\mathrm{Var}}(\bar{R}_1 - \bar{R}_2)} = \frac{\frac{n_1 n_2}{n_1+n_2}(\bar{R}_1 - \bar{R}_2)^2}{\sigma_R^2}, \tag{4.3}$$

where $\bar{R}_i = \sum_{k=1}^{n_i} R_{ik}/n_i$, n_i is the number of ranks in sequence $i, i = 1, 2$, $R_{ik}, k = 1, 2, \ldots, n_i$ are the ranks for that sequence and

$$\sigma_R^2 = \frac{\sum_{i=1}^{2}\sum_{k=1}^{n_i}(R_{ik} - \bar{R})^2}{(n_1 + n_2 - 1)} \quad .$$

On the null hypothesis that the distributions of T and R are equal, Q has an asymptotic chi-squared distribution on 1 degree of freedom.

Let $W_1 = \sum_{k=1}^{n_1} R_{1k}$ denote the rank-sum in the first sequence. We will now show that (4.3) is the square of the Wilcoxon rank-sum test statistic. The numerator of this statistic is

$$W_1 - E(W_1) = W_1 - \frac{n_1(n_2 + n_1 + 1)}{2} =$$

$$n_1\bar{R}_1 - \frac{n_1(n_2 + n_1 + 1)}{2} = n_1(\bar{R}_1 - \bar{R}),$$

where $\bar{R} = (\sum_{i=1}^{2}\sum_{k=1}^{n_i} R_{ik})/(n_1 + n_2) = (n_1 + n_2 + 1)/2$. In addition, as $\bar{R}_1 - \bar{R} = \bar{R}_1 - \frac{(n_1\bar{R}_1 + n_2\bar{R}_2)}{n_1+n_2}$, we have

$$W_1 - \frac{n_1(n_2 + n_1 + 1)}{2} = \frac{n_1 n_2(\bar{R}_1 - \bar{R}_2)}{n_1 + n_2} \quad .$$

In the absence of ties

$$\mathrm{Var}(W_1) = \frac{n_1 n_2(n_1 + n_2 + 1)}{12} \quad .$$

Returning now to Equation (4.3),

$$\sigma_R^2 = \frac{(n_1+n_2)(n_1+n_2+1)}{12} = \frac{n_1+n_2}{n_1 n_2}\,\mathrm{Var}(W_1) \quad .$$

Hence,

$$Q = \frac{[W_1 - E(W_1)]^2}{\mathrm{Var}(W_1)} \quad . \tag{4.4}$$

To illustrate this we consider the first stratum in Table 4.18 and first calculate the test statistic in more conventional ways. The calculations are done on the log-scale. Using StatXact, for example, and using asymptotic inference, the Wilcoxon rank-sum test statistic is -1.6922, which is asymptotically $N(0,1)$ on the null hypothesis. Using SAS `PROC FREQ` to calculate the corresponding Cochran-Mantel-Haenszel statistic with modified ridit scores, the test statistic is 2.8636 ($= 1.6922^2$), which is asymptotically chi-squared on 1 d.f. under the null hypothesis. The corresponding two-sided P-value is 0.0906.

To calculate Q, as defined in (4.4), we note that $\bar{R}_1 = 8.6$, $\bar{R}_2 = 13.1818$, $n_1 = 10$, $n_2 = 11$ and $\sigma_R^2 = 38.40$. Hence, $Q = [(10 \times 11)(8.60 - 13.1818)^2]/38.40 = 2.8636$. In the following we will use the form of the test-statistic defined in Equation (4.3).

Before moving on, it is useful to demonstrate one further way of calculating the numerator of the test statistic. Let d_{ik} denote the kth period difference in sequence $i, i = 1, 2$. The $n_1 n_2$ differences defined as $w_{\{k,k'\}} = d_{1k} - d_{2k'}$, where $k = 1, 2, \ldots, n_1$ and $k' = 1, 2, \ldots, n_2$, are known as the Walsh differences. Let s_j denote a weight for stratum j, where $j = 1, 2, 3$. For the moment we are dealing with only one stratum, so we set $s_1 = 1$. For comparison with later equations we will keep s_1 in the following formulae, even though it is unnecessary for the case of a single stratum. We will use d to denote the shift difference between the distributions of d_{1k} and $d_{2k'}$.

The rank sum for group i can be written as:

$$U_i = \sum_{\{w_{k,k'} > \, d\}} s_1 + 0.5 \sum_{\{w_{k,k'} = \, d\}} s_1 . \tag{4.5}$$

Further, as $U_1 + U_2 = n_1 n_2$ and $U_i = W_i - n_i(n_i+1)/2$,

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}(\bar{R}_1 - \bar{R}_2) = \frac{U_1 - U_2}{2},$$

where $\bar{R}_i = W_i/n_i$. Finally,

$$\frac{U_1 - U_2}{2} = \sum_{\{w_{k,k'} > d\}} s_1 + 0.5 \sum_{\{w_{k,k'} = \, d\}} s_1 - 0.5 n_1 n_2 s_1 .$$

The Wilcoxon rank-sum test statistic can then be expressed as:

$$W(d) = \frac{\sum_{\{w_{k,k'} > \ d\}} s_1 + 0.5 \sum_{\{w_{k,k'} = \ d\}} s_1 - 0.5 \ n_1 n_2 \ s_1}{\sqrt{\frac{n_1 n_2}{(n_1+n_2)}} \sigma_R}. \qquad (4.6)$$

Rearranging Equation (4.6) gives

$$W(d)\sqrt{\frac{n_1 n_2}{(n_1+n_2)}}\sigma_R + 0.5 n_1 n_2 s_1 = \sum_{\{w_{k,k'} > \ d\}} s_1 + 0.5 \sum_{\{w_{k,k'} = \ d\}} s_1. \qquad (4.7)$$

Solving Equation (4.7) with $W(d) = 0$ gives the median of the Walsh differences (-0.293), and this is (twice) the estimate of δ.

Solving Equation (4.7) with $W(\delta) = \pm 1.645$ gives the positions of the Walsh differences that correspond to (twice) the lower and upper 90% confidence bounds for δ.

For the first stratum,

$$1.645\sqrt{\frac{n_1 n_2}{(n_1+n_2)}}\sigma_R + 0.5 n_1 n_2 s_1 = 23.33 + 55 = 78.3.$$

The 79th value in the ordered set of Walsh differences (not shown) is 0.0. For the lower bound we take the $-23.33 + 55 = 31.67$, i.e., the 31st ordered difference which is -0.629. The 90% confidence interval for δ is therefore $(-0.314, 0.000)$. If we take the limits for bioequivalence to be $(-0.223, 0.223)$ as for AUC and Cmax, then there is clear evidence that T and R are not bioequivalent when Tmax is used as the metric.

For the remaining two strata the estimate and confidence intervals for δ are, respectively, [-0.463, -0.247, -0.098] and [-0.289, -0.143, 0.007]. Again there is strong evidence of a lack of equivalence.

Multiple Strata

The extension of the Wilcoxon rank-sum test statistic to multiple strata is:

$$W = \frac{\sum_{i=1}^{q} s_i W_i}{\sqrt{\sum_{i=1}^{q} s_i^2 \operatorname{Var}(W_i)}} = \frac{\sum_{i=1}^{q} s_i W_i}{\sqrt{\sum_{i=1}^{q} s_i^2 \frac{n_{i1} n_{i2}}{n_{i1}+n_{i2}} \sigma_{iR}^2}}, \qquad (4.8)$$

where W_i is the rank-sum statistic for the ith single stratum, $\operatorname{Var}(W_i)$ is its variance, s_i is the weight for the ith stratum, σ_{iR}^2 is the variance of the ranks in the ith stratum, and n_{ij} is the number of ranks in sequence group j in stratum i. We will use the weights suggested by Lehmann [261]: $s_i = 1/(n_{i1} + n_{i2} + 1)$, which give the Van-Elteren test statistic. However, for our purposes we require the corresponding 90% confidence interval. In a way similar to that described for a single stratum we can

Table 4.20 *Period Differences for Comparing T and R: Stratum I*

Subject	Difference	Difference (log scale)	Stratum	Group	Rank
6	0.00	0.000	I	1	11
12	-0.02	-0.005	I	1	9
17	-1.00	-0.289	I	1	4
19	2.48	0.976	I	1	21
29	-0.96	-0.276	I	1	7
32	0.02	0.010	I	1	14
42	-1.00	-0.288	I	1	5
46	-0.98	-0.283	I	1	6
49	-0.52	-0.233	I	1	8
60	-2.48	-0.976	I	1	1
2	0.00	0.000	I	2	11
11	0.97	0.395	I	2	17
16	1.00	0.288	I	2	16
20	0.98	0.396	I	2	18
27	-1.98	-0.688	I	2	2
31	0.98	0.503	I	2	19
40	0.47	0.273	I	2	15
43	0.04	0.010	I	2	13
53	-1.48	-0.465	I	2	3
59	0.00	0.000	I	2	11
61	2.00	0.693	I	2	20

write the numerator of (4.8) as:

$$\sum_{i=1}^{q} \sum_{\{w_{ik,ik'} > \, d\}} s_i + 0.5 \sum_{i=1}^{q} \sum_{\{w_{ik,ik'} = \, d\}} s_i - 0.5 \sum_{i=1}^{q} n_{i1} n_{i2} s_i. \qquad (4.9)$$

Rearranging Equation (4.8), we get

$$\begin{aligned} W(d)\sqrt{\sum_{i=1}^{q} s_i^2 \frac{n_{i1} n_{i2}}{n_{i1}+n_{i2}} \sigma_{iR}^2} + 0.5 \sum_{i=1}^{q} n_{i1} n_{i2} s_i \\ = \sum_{i=1}^{q} \sum_{\{w_{ik,ik'} > \, d\}} s_i + 0.5 \sum_{i=1}^{q} \sum_{\{w_{ik,ik'} = \, d\}} s_i \ . \qquad (4.10) \end{aligned}$$

As before, we set $W(d) = 0$ and solve to get the estimator of $2d$. Setting $W(d) = \pm 1.645$ gives the positions of the Walsh differences that correspond to (twice) the lower and upper 90% confidence bounds for

Table 4.21 *Period Differences for Comparing T and R: Stratum II*

Subject	Difference	Difference (log scale)	Stratum	Group	Rank
4	-0.52	-0.190	II	1	6.0
7	-1.49	-0.470	II	1	4.0
14	-0.02	-0.007	II	1	9.5
23	-2.00	-1.099	II	1	1.0
26	-1.95	-0.393	II	1	5.0
36	-0.02	-0.007	II	1	9.5
39	0.10	0.025	II	1	13.0
48	-0.99	-0.674	II	1	2.0
54	-0.02	-0.008	II	1	8.0
56	-0.98	-0.503	II	1	3.0
9	0.48	0.176	II	2	14.0
13	0.50	0.288	II	2	16.0
21	0.97	0.486	II	2	17.0
28	-0.47	-0.172	II	2	7.0
33	1.95	1.069	II	2	18.0
44	0.03	0.008	II	2	12.0
50	-0.02	-0.005	II	2	11.0
58	0.52	0.190	II	2	15.0

δ. For T versus R, $\hat{\delta} = 0.192$ with confidence interval (0.136, 0.279) and for T versus S $\hat{\delta} = 0.054$ with confidence interval (-0.005, 0.145).

In summary, there is evidence is that T and R are not equivalent but T is equivalent to S.

Bootstrap estimation of confidence intervals

An alternative method of getting a nonparametric estimate of the 90% confidence interval for $\mu_T - \mu_R$ is to use bootstrapping. (See Chapter 5 for a more detailed explanation of the bootstrap.) The method as applied here is to resample with replication from the 60 sets of triples (the three repeated measurements on each subject) and to calculate an estimate of $\mu_T - \mu_R$ from each resample. If this is done a large number of times, say 1000 times, a distribution of the estimator is generated. The 5% and 95% quantiles of this distribution provide a 90% confidence interval for $\mu_T - \mu_R$. The median of this distribution is an estimate of $\mu_T - \mu_R$. There will usually be a choice of estimator to use. Here we have taken the least squares estimator obtained by fitting a linear model with terms for subjects, period, and treatments. The distributions for $\mu_T - \mu_R$ and $\mu_T - \mu_S$ obtained from 1000 resamples are given in Figure 4.20. The quantiles and medians obtained are: (0.0889, 0.1811, 0.2670) for $\mu_T - \mu_R$

Table 4.22 *Period Differences for Comparing T and R: Stratum III*

Subject	Difference	Difference (log scale)	Stratum	Group	Rank
5	-0.57	-0.253	III	1	9
10	-1.00	-0.511	III	1	3
18	0.00	0.000	III	1	11
22	-1.00	-0.286	III	1	6
44	-0.96	-0.276	III	1	7
34	0.02	0.005	III	1	12
37	-2.54	-0.999	III	1	1
47	-1.00	-0.288	III	1	5
52	1.48	0.462	III	1	17
55	1.50	0.473	III	1	20
1	1.02	0.293	III	2	16
8	-2.02	-0.703	III	2	2
15	-1.48	-0.465	III	2	4
24	-0.02	-0.005	III	2	10
25	0.02	0.007	III	2	13
35	1.48	0.465	III	2	18
41	-0.93	-0.266	III	2	8
45	1.00	0.285	III	2	15
51	3.48	0.867	III	2	21
57	0.98	0.283	III	2	14
62	1.50	0.470	III	2	19

and (-0.0363, 0.0481, 0.1253) for $\mu_T - \mu_R$. The conclusions obtained are consistent with those obtained from the nonparametric method. The only difference of note is that the lower limits of the bootstrap confidence intervals differ a little from those obtained earlier.

4.7.2 Four Treatments

The data in Table 4.39 are the Tmax values obtained in the trial described in Example 4.6. The comparisons of interest were B versus A and D versus C. It is clear from the design of this trial that the sequences cannot be grouped in a way that would allow the nonparametric approach described in the last subsection to be applied. However, we can use the bootstrapping approach. The 90% confidence intervals and medians so obtained are: (-0.2724, -0.0628, 0.1604) for $\mu_B - \mu_A$ and (-0.1062, 0.0703, 0.2554) for $\mu_D - \mu_C$. There is clear evidence of a lack of equivalence for both sets of treatments.

Table 4.23 *Components of the Stratified Test for Comparing T and R*

Statistic	Stratum I	Stratum II	Stratum III
Patients in Each Sequence	(10,11)	(10,8)	(10,11)
Rank Sum in each Sequence(W)	(86, 145)	(61, 110)	(91, 140)
$E(W)$	(110,121)	(95,76)	(110,121)
$W - E(W)$	(-24,24)	(-34,34)	(-19,19)
Estimated Variance of Test Statistic	38.40	28.471	38.50
Weight	0.045	0.053	0.045
Rank-Sum Statistic	-1.692	-3.022	-1.338

4.8 Technical Appendix: Efficiency

4.8.1 Theory and Illustration

We assume that our model for the response includes terms for a general mean, fixed subject effects, periods, formulations, and carry-over effects. Let the responses, e.g., logAUC, be stored in a random vector $\mathbf{y}$ which is assumed to have mean vector $\mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2_W\mathbf{I}$. Here $\mathbf{X}$ is a design matrix with elements that are either 0 or 1, $\boldsymbol{\beta}$ is a vector of unknown subject, period, formulation, and carry-over parameters and $\mathbf{I}$ is the identity matrix with row and column dimension equal to that of $\mathbf{y}$. The parameters are estimated by ordinary least squares:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

with

$$\mathrm{V}(\hat{\boldsymbol{\beta}}) = \sigma^2_W(\mathbf{X}^T\mathbf{X})^{-1}.$$

We assume that any redundant parameters have been removed and $\mathbf{X}^T\mathbf{X}$ is of full rank. This can be achieved, for example, by removing one subject parameter, one period parameter, one formulation parameter and one carry-over parameter. If the design is for n subjects with n of them randomly allocated to each of the two sequences RTT and TRR, there

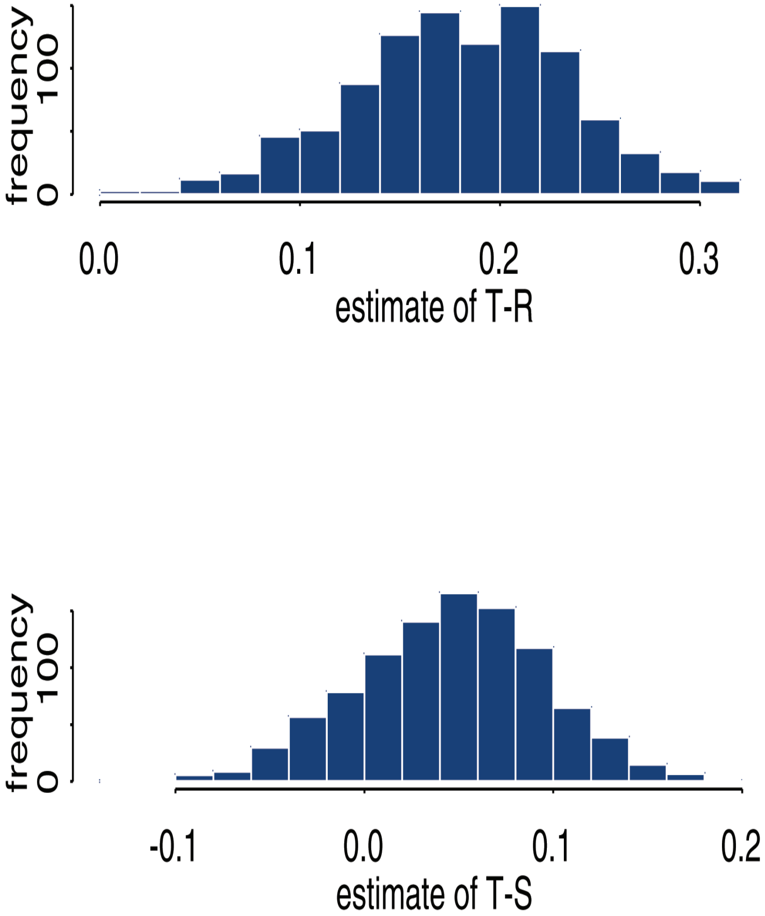

Figure 4.20 *Example 4.5: Histograms of Bootstrap Distribution of Estimates*

will be $(1 + n - 1 + 1 + 1 + 1 = n + 3)$ parameters. However, we do not need to work with this many parameters to calculate the efficiency. Jones and Kenward [237] show that this can be done using the corresponding design with one subject allocated to each sequence. In other words, we put parameters in the model for sequences instead of subjects. We will illustrate this in the following.

The efficiency of a design compares (1) the variance of the estimated difference between two formulations in the given design to (2) the corresponding variance in an ideal design with the same formulation replication and same within-subject variance σ^2_W. The ideal design is such that it would completely eliminate the effects of subjects, periods, and carry-over effects from the estimation of the formulation comparison. For example, suppose that T and R each occur r times in the ideal design. The estimate of the formulation difference is $\bar{y}_T - \bar{y}_R$ and its variance is $V_I = 2\sigma^2_W/r$. This is used as the benchmark for other designs.

For the particular cross-over design under consideration, e.g., one with sequences RTT and TRR, and using the particular parametrization given above, the treatment parameter, τ_2 corresponds to the difference between T and R. The variance of this difference is the diagonal element of $\sigma^2_W(X^TX)^{-1}$ that occurs in the position corresponding to τ_2 in the vector of parameters. We will give an example of locating this element below. Let us call this element $V_C = \sigma^2 a_C$.

The efficiency of the cross-over design for the T-R difference is then the percentage:

$$E = 100 \times \frac{V_I}{V_C} = 100 \times \frac{2}{r \times a_C}.$$

Efficiency cannot exceed 100%.

As an example consider the design with sequences RTT and TRR and $n/2$ subjects per sequence. Suppose we want to allow for a difference in carry-over effects and put these into our model. For the basic calculations we assume $n = 2$, then scale down the variances and covariances according to the true value of n. The design matrix for the model with sequence, period, formulation, and carry-over effects is as follows, where redundant parameters have been removed:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The columns in this matrix refer to the general mean, Sequence 2, Periods 2 and 3, Formulation T, and the carry-over of T, respectively. Although there is no carry-over effect in the first period, we must include a 'dummy' parameter to represent this missing effect if we are to construct the $\mathbf{X}$ matrix. Our way of doing this is to let the carry-over parameter for T do 'double-duty' by also taking on the role of this dummy parameter. As long as there are period effects in the model, there is no confusion because the dummy parameter is aliased with parameter for Period 1 and effectively gets removed correctly in the analysis. The inverse matrix, from which the variances are taken or calculated, is:

$$\sigma_W^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma_W^2}{4}\begin{bmatrix} 4 & -2 & 2 & -2 & -2 & 0 \\ -2 & 3 & 0 & 0 & 1 & 0 \\ -2 & 0 & 5 & 3 & 0 & -2 \\ -2 & 0 & 3 & 5 & 0 & -2 \\ -2 & 1 & 0 & 0 & 3 & 0 \\ 0 & 0 & -2 & -2 & 0 & 4 \end{bmatrix}.$$

This inverse is for a design with one subject per sequence. To get the correct value of a variance of a comparison of means we divide the elements of this inverse by the number of responses used in calculating the means. For example, when there are $n/2$ subjects per sequence the variance of the estimate of T-R, adjusted for carry-over, is $(3\sigma_W^2/4)/(n/2)$, i.e., $a_C = (3/4)/(n/2) = 3/(2n)$ and the variance of the corresponding estimated carry-over difference is $\sigma_W^2/(n/2)$, i.e., $a_C = 2/n$. The required elements of $\sigma_W^2(\mathbf{X}^T\mathbf{X})^{-1}$ are those in the fifth and sixth positions along the diagonal because the parameters that refer to T-R and the carry-over difference are in these positions, respectively, in the vector $\boldsymbol{\beta}$. Because the (5,6)th element of $\sigma_W^2(\mathbf{X}^T\mathbf{X})^{-1}$ is zero, these two estimates are uncorrelated. We are now in a position to calculate the efficiency of the T-R comparison. As each formulation occurs $3n/2$ times in the design, $V_I = 4/3n$ and hence:

$$E = 100 \times \frac{2}{r \times a_C} = 100 \times \frac{2}{(3n/2)(3/2n)} = 100 \times \frac{8}{9} = 88.9\%.$$

Although we are not usually interested in the efficiency of the carry-over comparison, we will calculate it for completeness and as a further illustration. Traditionally, the replication for each carry-over effect is taken to be that of the corresponding formulation, e.g., $3n/2$ in the above design. However, as there are no carry-over effects in the first period, this replication is strictly too large. However, we will stick with the traditional approach. Hence, the efficiency of the comparison of the carry-over effects of T and R is:

$$E = 100 \times \frac{2}{r \times a_C} = 100 \times \frac{2}{(3n/2)(2/n)} = 100 \times \frac{2}{3} = 66.7$$

We note that the efficiencies for an arbitrary cross-over design can be calculated using the GenStat statistical analysis system via the procedure XOEFFICIENCY [239].

4.8.2 Comparison of Three Alternative Designs for Three Periods

Here we compare three alternative designs that could be used to compare T and R in a bioequivalence trial. These are listed below:

1.	R	T	T	2.	R	T	R	3.	R	R	T
	T	R	R		T	R	T		T	T	R.

The efficiencies of the formulation and carry-over comparisons are given in Table 4.24, where we have also included the correlation between the estimators of the formulation and carry-over differences. A major advantage of the first design is that the estimator of the formulation difference does not change if the carry-over parameter is left out of the model, as the correlation is zero. Hence, this design provides some robustness against the presence of a carry-over difference, which, although unexpected, cannot always be ruled out entirely.

Table 4.24 *Efficiencies of Three Alternative Designs*

Design	Formulation T-R	Carry-over T-R	Correlation (Formulation, Carry-over)
1. RTT/TRR	88.9	66.7	0.00
2. RTR/TRT	22.2	16.7	0.87
3. RRT/TTR	66.7	16.7	0.50

4.9 Tables of Data

Table 4.25: Example 4.1: Sequence RTT

	Sequence RTT					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
104	37.27	62.18	44.09	2.207	2.901	2.073
105	82.870	24.780	24.700	6.123	1.462	1.468
106	47.800	32.880	124.310	2.586	1.203	6.972
107	88.390	30.850	192.450	4.326	1.589	8.687
108	180.50	108.71	200.57	8.459	5.011	9.104
111	50.59	33.53	100.58	3.133	1.814	7.159
113	634.140	914.900	-	7.154	12.354	8.207
115	420.300	205.740	-	20.221	11.746	-
117	582.260	736.820	784.960	9.819	12.035	17.973
118	45.420	-	70.690	1.636	0.852	1.895
R=Reference, T=Test						

Table 4.25: Example 4.1: Sequence RTT

	Sequence RTT					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
120	437.610	586.470	405.950	9.111	11.708	10.539
121	22.830	13.720	15.750	1.167	0.506	0.756
123	64.58	35.54	65.11	2.949	1.831	2.989
126	15.15	22.35	21.71	0.902	1.234	1.495
128	30.220	27.400	33.190	1.632	0.921	1.221
130	12.420	71.380	62.270	0.636	4.433	4.408
133	39.010	89.410	59.890	1.854	4.091	2.235
136	24.470	42.660	42.390	1.441	2.997	3.070
137	13.840	21.730	41.690	0.846	1.202	2.380
138	28.040	10.970	42.720	1.045	0.629	2.337
139	264.890	243.660	276.540	13.913	9.160	10.632
141	-	-	-	0.355	0.237	0.444
142	227.010	8.080	521.640	11.638	0.655	23.115
147	71.100	16.770	44.080	3.489	1.013	2.434
150	29.660	76.030	60.120	1.439	5.327	4.626
153	1737.430	1416.780	1336.790	21.715	22.405	16.726
154	440.830	163.920	282.290	25.232	6.205	11.416
155	53.830	48.090	78.280	1.715	1.239	2.470
160	41.580	259.550	113.840	2.087	11.067	4.379
161	327.530	210.820	453.230	6.741	3.742	10.083
162	45.570	30.130	83.960	1.876	1.230	6.274
164	142.000	146.630	124.380	5.982	5.288	5.456
168	15.230	31.890	71.680	1.020	1.459	4.637
170	76.490	82.700	114.290	4.224	4.131	6.619
173	87.330	51.370	96.460	5.726	2.431	4.939
174	787.890	737.740	338.520	31.224	23.271	12.711
175	1239.480	1819.440	2232.290	24.013	30.484	43.224
177	29.190	36.580	79.590	1.971	2.296	4.243
179	10.130	16.990	9.820	1.029	1.371	0.718
181	257.590	423.890	224.070	9.964	15.005	6.776
182	51.770	27.630	26.090	3.797	2.312	1.741
184	73.750	90.810	-	2.555	3.242	-
185	49.320	124.000	85.710	1.471	4.079	4.743
186	6.060	28.820	87.630	0.311	1.651	4.870
190	82.780	164.560	213.980	3.889	7.376	7.012
191	98.860	99.020	75.480	4.599	2.969	2.388
R=Reference, T=Test						

Table 4.25: Example 4.1: Sequence RTT

	Sequence RTT					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
194	21.290	46.300	15.410	1.513	2.741	1.411
R=Reference, T=Test						

Table 4.26: Example 4.1: Sequence TRR

	Sequence TRR					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
101	12.260	16.190	11.340	0.511	0.688	0.533
102	397.980	267.630	487.550	13.270	7.933	12.952
103	243.810	141.700	198.440	16.771	6.926	9.257
109	182.520	112.340	225.940	8.816	4.921	6.911
110	559.640	533.980	867.750	21.398	19.728	19.909
112	40.020	89.490	20.350	2.568	5.222	0.992
114	-	-	34.810	0.872	0.337	1.558
116	69.380	214.200	193.620	3.543	8.911	5.900
119	68.080	47.190	84.590	2.673	1.501	4.187
122	181.950	259.400	396.260	5.841	10.642	19.245
124	5.820	17.260	25.720	0.347	1.241	1.175
125	39.310	35.660	40.430	2.288	1.786	2.589
127	146.870	319.910	141.860	5.772	10.780	6.768
129	712.110	549.520	459.260	16.116	13.171	10.648
131	2277.520	3726.580	3808.790	18.448	34.145	41.876
132	1278.060	1103.460	1012.040	18.779	17.086	13.170
134	103.320	138.780	170.440	4.974	5.349	8.128
135	21.930	75.290	42.300	1.622	4.791	3.228
140	77.990	104.080	66.860	3.043	5.210	2.625
143	27.210	47.190	25.340	1.170	2.405	1.698
144	296.090	163.310	387.490	10.730	6.443	13.790
145	82.600	247.710	92.940	3.363	9.128	5.311
146	18.010	241.700	205.390	1.011	10.183	9.865
148	123.270	268.090	128.170	4.985	8.893	5.880
149	52.460	201.680	421.550	2.457	6.945	32.983
151	29.830	20.660	24.550	1.691	1.186	1.313
R=Reference, T=Test						

Table 4.26: Example 4.1: Sequence TRR

	Sequence TRR					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
152	414.990	247.580	419.530	14.735	9.851	12.724
156	213.240	87.550	178.660	7.510	2.793	5.323
157	13.580	7.160	10.940	0.496	0.459	0.756
158	172.250	211.290	206.990	7.330	5.667	9.804
159	1161.730	2280.790	1552.490	27.604	45.495	27.220
163	57.260	48.650	89.010	2.691	2.877	6.631
165	350.950	755.270	711.180	7.034	13.040	11.002
166	36.79	41.75	35.39	1.861	2.75	2.784
167	11.57	3.31	-	1.055	0.326	0.296
171	28.440	61.400	25.500	1.246	3.146	1.016
172	1150.280	759.030	1105.080	15.677	15.215	20.192
176	69.630	24.020	26.110	3.971	1.234	0.948
178	179.76	190.89	299.5	4.909	5.374	10.014
180	14.23	22.44	23.70	1.088	1.783	1.733
183	295.690	304.030	277.670	11.125	9.916	10.649
187	34.180	45.140	58.670	1.870	3.055	4.654
188	50.380	87.620	16.460	2.317	4.658	0.719
189	104.08	123.08	129.00	3.73	4.109	6.018
192	17.19	40.01	55.36	1.994	2.786	3.716
193	131.570	156.120	130.480	7.191	12.207	7.532
195	1323.070	1305.500	2464.820	12.897	24.767	27.650
196	654.320	783.530	444.440	12.347	26.041	18.975
R=Reference, T=Test						

Table 4.27: Example 4.2: Sequence RTT

	Sequence RTT					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
1	1158.06	1073.74	748.58	15.44	11.93	14.12
4	520.75	410.53	437.96	13.59	9.17	8.85
5	11.44	13.29	14.31	0.70	0.80	0.92
6	-	28.87	19.44	0.68	1.19	1.44
9	51.76	23.75	35.23	2.48	1.20	1.97
R=Reference, T=Test						

Table 4.27: Example 4.2: Sequence RTT

	Sequence RTT					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
10	-	8.93	5.85	0.35	0.79	0.46
15	25.80	27.91	51.47	1.42	1.78	3.24
16	1633.77	1127.82	1267.52	20.18	35.76	16.24
18	105.03	15.61	18.03	5.87	0.81	0.93
19	1635.06	1562.78	1936.28	20.91	18.53	17.17
22	168.29	337.16	227.49	5.82	10.45	5.45
23	3.23	7.84	4.86	0.28	0.64	0.54
25	44.81	12.22	24.56	2.73	0.78	1.53
28	15.54	24.71	29.74	0.91	1.01	1.33
29	48.69	17.61	35.34	3.66	1.22	1.71
32	134.01	204.85	81.73	5.26	7.51	2.91
34	48.15	17.59	20.08	3.60	1.21	1.15
35	39.22	13.58	19.21	5.27	0.99	1.57
36	805.16	602.79	698.12	20.15	12.13	13.05
37	52.97	55.85	44.97	3.46	4.31	2.70
38	23.07	-	39.34	1.02	2.09	1.31
42	46.99	59.85	60.41	2.33	3.54	2.90
47	43.37	50.40	85.98	2.06	2.73	4.02
48	12.25	9.59	11.70	0.72	0.80	0.39
49	15.47	13.90	19.09	0.80	1.04	0.94
50	54.21	93.00	121.17	1.71	3.90	4.77
53	38.92	32.07	61.57	2.78	1.94	3.05
55	947.92	707.40	696.01	11.72	9.97	9.34
57	37.40	78.42	85.38	1.91	4.13	3.55
62	64.95	66.42	91.42	2.74	3.78	5.06
63	9.38	10.95	18.37	1.16	0.77	1.32
67	132.73	128.11	135.28	10.58	5.92	5.56
68	140.46	97.09	153.54	8.52	6.03	7.50
70	366.38	300.67	275.54	13.50	13.41	11.15
71	48.65	40.87	-	2.96	3.08	3.02
73	544.33	617.22	554.04	11.07	13.69	13.11
75	16.69	9.65	13.68	1.90	0.57	1.16
79	60.85	41.24	39.05	2.25	1.76	2.91
80	38.90	61.10	40.88	2.24	3.68	2.50
R=Reference, T=Test						

Table 4.28: Example 4.2: Sequence TRR

	Sequence TRR					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
2	17.28	30.30	83.53	1.20	2.23	5.25
3	11.63	16.20	18.23	0.75	1.34	1.27
7	78.03	42.64	148.29	3.80	1.28	5.11
8	6.61	19.83	7.18	0.64	1.22	1.06
11	14.68	16.74	25.73	1.06	1.74	2.89
12	119.77	211.51	148.04	5.07	9.11	4.78
13	36.26	34.02	50.11	2.59	2.29	2.93
14	59.06	94.61	54.46	4.84	5.79	3.03
17	17.47	39.47	31.08	1.41	2.94	2.49
20	1082.90	1497.28	2011.67	21.62	29.04	29.89
24	47.84	46.22	68.04	3.10	3.16	4.48
26	-	19.24	20.01	0.59	1.08	1.54
27	26.30	15.45	88.92	2.15	1.20	4.78
30	23.94	54.15	55.25	1.47	3.07	2.09
31	21.90	18.72	15.20	1.02	1.08	1.02
33	20.20	28.40	44.84	1.52	1.44	2.59
39	59.06	87.12	148.31	2.93	3.50	6.57
40	79.04	31.79	64.29	4.87	1.65	2.93
41	139.30	74.26	92.94	6.96	4.53	5.36
43	503.28	389.44	547.82	10.86	9.53	10.44
45	50.24	52.74	57.02	2.15	2.66	2.32
46	29.35	41.32	33.12	2.02	2.14	1.79
51	-	20.66	8.13	1.25	2.67	0.53
52	26.95	50.10	26.56	1.67	2.74	1.37
54	19.48	12.62	18.78	1.32	0.64	1.30
56	20.27	-	10.64	1.71	0.65	0.94
61	14.57	49.60	58.36	1.06	2.34	2.97
64	56.74	61.83	97.05	3.62	3.12	4.82
65	103.19	187.82	188.43	5.65	8.45	8.41
69	13.12	32.13	18.02	0.94	2.11	0.99
72	14.90	16.00	11.85	1.17	0.94	0.66
74	24.60	39.14	53.98	1.31	2.42	3.63
76	7.50	4.80	12.06	0.52	0.44	1.50
77	828.00	565.73	1085.51	13.37	7.32	14.84
78	33.99	47.96	35.15	2.65	3.17	2.04

R=Reference, T=Test

Table 4.29: Example 4.3

	Sequence RTTR							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
1	10671	12772	13151	11206	817	1439	1310	1502
4	7588	8980	8408	7654	823	1133	1065	1095
6	8389	7949	7735	7616	1347	691	949	1153
7	5161	6601	5479	4764	1278	991	1124	1040
9	7399	7873	8153	7211	1547	1361	1380	1485
10	5660	4858	5347	5076	1088	982	995	796
15	6937	7905	6550	7515	953	1065	830	1247
16	11473	9698	10355	10365	1368	1281	1083	1418
	Sequence TRRT							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
2	6518	6068	5996	5844	1393	1372	1056	1310
3	4939	5728	5760	6313	1481	1377	1529	781
5	7653	8022	10721	8043	709	1035	1571	1342
8	8864	8026	6776	6995	1516	1242	1090	1048
11	8503	7730	8228	8032	999	908	1183	1129
12	7043	6007	7737	6262	679	1220	776	1258
13	5701	5767	5942	7757	822	869	921	947
14	8684	7858	7924	9219	615	1451	1389	1279
18	5210	5120	5420	-	668	842	1176	-
R=Reference, T=Test								

Table 4.30: Example 4.4

	Sequence RTRT							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
1	812.60	1173.70	889.10	620.10	99.85	204.09	170.94	112.78
3	545.10	542.90	-	-	67.69	41.73	-	-
5	400.00	223.80	173.70	289.70	40.05	25.17	24.48	86.49
6	102.10	185.30	42.00	88.30	28.76	24.83	9.27	10.89
10	304.50	351.50	520.20	335.70	34.35	52.26	142.92	58.48
12	176.10	710.70	409.50	645.50	18.94	161.34	118.89	246.57
15	562.40	490.40	504.70	675.90	28.35	98.50	78.22	140.54
17	207.50	271.60	173.70	240.50	19.18	94.92	21.39	65.45
18	571.30	705.20	619.00	633.60	66.63	134.69	78.10	78.51
21	536.10	595.20	445.50	521.50	42.11	37.82	39.87	116.79
24	449.90	860.40	606.80	577.20	32.53	276.86	118.65	156.33
25	192.50	220.10	233.10	227.00	21.96	38.97	22.26	54.16
28	568.10	321.10	338.30	403.60	110.87	55.64	50.06	84.60
29	735.60	634.50	1244.20	641.90	50.08	58.79	181.53	144.26
31	307.40	481.80	346.60	369.70	87.21	88.75	90.07	132.92
34	292.90	431.00	448.50	267.80	18.07	33.37	21.48	20.87
35	217.20	332.20	103.00	127.50	18.69	174.55	17.06	32.01
39	368.30	292.60	446.10	222.30	52.59	57.88	48.58	47.24
40	193.70	202.80	255.20	244.30	29.30	78.33	21.72	49.27
R=Reference, T=Test								

Table 4.30: Example 4.4

	Sequence RTRT							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
44	102.00	282.50	245.60	286.20	22.14	63.50	9.38	16.30
46	223.60	645.40	349.00	507.40	27.02	167.28	20.35	121.92
48	615.80	732.10	620.90	665.20	60.94	100.47	26.17	98.08
49	898.40	924.90	398.30	828.30	164.01	180.01	25.21	97.02
50	410.40	329.20	449.40	442.10	59.70	43.65	102.47	40.00
53	332.40	273.60	525.30	293.30	39.96	56.47	42.11	38.75
54	185.20	222.90	182.10	194.10	18.34	16.09	21.50	9.57
57	180.60	174.70	102.90	117.00	9.10	58.44	12.74	18.33
R=Reference, T=Test								

Table 4.31: Example 4.4

	Sequence TRTR							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
Sub	1	2	3	4	1	2	3	4
2	216.30	338.00	502.80	398.60	29.06	50.48	35.15	55.71
4	632.60	520.00	716.70	860.40	91.25	43.86	168.78	61.04
7	596.00	659.30	543.80	662.90	257.10	79.04	127.92	81.80
8	402.40	359.80	590.80	444.30	136.27	158.86	148.97	82.41
9	456.70	378.40	477.50	407.90	65.48	87.84	64.57	58.01
11	500.70	323.00	416.30	525.10	31.49	37.07	80.90	33.62
13	160.60	218.00	170.10	124.60	29.61	43.15	27.71	13.11
16	756.00	606.80	477.40	626.80	168.76	174.94	117.31	52.18
19	511.90	549.70	388.20	141.00	32.23	70.06	32.15	43.11
20	124.00	91.90	113.30	59.50	9.34	11.74	49.23	18.42
22	239.70	265.10	445.90	433.20	38.02	16.79	38.58	83.82
23	609.60	371.60	511.30	432.70	199.07	52.14	118.47	72.04
26	764.40	508.80	757.80	449.40	74.24	35.76	39.27	36.28
27	151.90	194.80	-	-	19.00	20.61	-	-
30	429.10	391.80	316.90	335.10	31.85	74.88	54.88	19.18
32	409.00	514.60	763.10	406.50	30.86	70.84	208.20	65.25
33	271.00	221.00	296.50	463.70	86.01	41.85	67.86	79.81
36	290.80	208.60	243.70	489.80	38.27	40.31	31.56	20.64
37	297.20	502.00	320.40	334.30	49.81	66.64	17.80	25.94
38	163.80	232.10	636.90	434.90	34.56	16.37	114.30	29.58
42	534.10	243.10	418.40	441.90	136.00	33.75	104.12	35.03
43	355.10	415.20	382.70	334.00	64.55	34.04	52.37	41.67
45	320.50	233.90	331.70	260.50	26.35	37.20	76.26	24.60
47	504.50	289.90	550.70	244.20	118.91	49.27	166.61	35.86
52	237.00	505.00	496.30	580.60	30.55	63.90	39.17	40.75
55	246.90	620.90	678.30	752.20	42.20	106.69	150.52	115.15
56	235.40	190.40	318.30	248.40	39.15	13.79	122.03	62.32
R=Reference, T=Test								

Table 4.32: Example 4.5: Williams Design for 3 Treatments

	Sequence RST					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
9	4089	7411	5513	906	1711	1510
13	2077	3684	2920	504	845	930
21	2665	3113	2263	506	809	543
28	3029	5157	4190	563	1263	759
33	4941	4502	3014	1095	1253	1015
44	2173	4571	3350	366	1341	779
50	-	-	3900	602	1291	1314
R=3 × 100mg, S=200mg + 100mg, T=Test						

Table 4.32: Example 4.5: Williams Design for 3 Treatments

	Sequence RST					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
58	6555	11351	8895	1229	2138	2144
67	4045	7865	-	1025	2668	-
	Sequence RTS					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
2	3457	6556	4081	776	2387	1355
11	5560	4558	4396	1801	1440	1327
16	3676	5385	5358	544	1556	1776
20	8636	9750	9892	2238	2256	2277
27	2753	2736	3955	572	593	1142
31	4782	4812	4024	1078	1224	1010
40	2636	2791	2394	546	587	442
43	3011	4544	6587	558	998	1418
53	2685	5335	7454	530	1160	1764
59	4841	5934	6624	1416	1302	1517
61	2392	2947	3779	644	744	1144
R=3 × 100mg, S=200mg + 100mg, T=Test						

Table 4.33: Example 4.5: Williams Design for 3 Treatments

	Sequence SRT					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
1	7260	6463	8759	1633	1366	2141
8	3504	3011	2501	959	557	697
15	6641	1987	3233	1586	364	633
24	4368	4327	2966	991	748	1001
25	8016	7146	9154	2045	1891	2545
35	7749	4188	3425	1855	757	758
41	8961	8737	11312	1722	1313	2705
45	4537	2633	3723	999	604	1075
51	5658	4904	5077	1539	1227	1490
57	5194	2432	4472	1810	686	1149
R=3 × 100mg, S=200mg + 100mg, T=Test						

Table 4.33: Example 4.5: Williams Design for 3 Treatments

	Sequence SRT					
	AUC			Cmax		
	Period			Period		
Subject	1	2	3	1	2	3
62	5787	7069	6530	1461	1995	1236
	Sequence STR					
	AUC			Cmax		
	Period			Period		
Subject	1	2	3	1	2	3
5	4250	3487	2891	945	1041	788
10	4839	3064	2582	1051	991	782
18	6317	5175	3123	1432	1184	647
22	3527	3484	2580	656	734	531
30	2717	2743	1625	637	760	463
34	4709	3212	3840	1022	661	609
37	5256	4070	2505	1194	974	432
47	5840	5213	5213	1329	1477	1039
52	4622	2889	2692	1027	562	422
55	8671	6814	4260	2251	1561	1045
R=3 × 100mg, S=200mg + 100mg, T=Test						

Table 4.34: Example 4.5: Williams Design for 3 Treatments

	Sequence TRS					
	AUC			Cmax		
	Period			Period		
Subject	1	2	3	1	2	3
6	6709	5893	5346	1292	1154	1098
12	7026	6134	9520	1417	1207	2312
17	9249	5535	9965	2232	913	2887
19	4664	2998	6592	1103	547	2113
29	5547	7319	8331	1288	1506	1884
32	3500	5611	5394	852	1259	1308
42	4367	5827	8863	736	1135	2288
46	3020	3989	3739	643	660	841
49	-	-	6092	1556	1895	1854
60	3125	4728	3199	594	1317	731
63	2204	2927	-	495	770	-
	Sequence TSR					
R=3 × 100mg, S=200mg + 100mg, T=Test						

Table 4.34: Example 4.5: Williams Design for 3 Treatments

	Sequence TRS					
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
	AUC Period			Cmax Period		
Subject	1	2	3	1	2	3
4	4006	4879	3817	1326	1028	1052
7	6924	4674	4183	1475	994	1142
14	6027	6497	5048	1106	1914	1358
23	2642	3178	2496	461	589	561
26	3064	3534	2302	754	1508	419
36	9882	13881	6881	2054	3042	1207
39	1422	2375	1559	316	555	427
48	6029	4114	3625	2261	1097	1038
54	5429	7513	4589	1369	2068	1384
56	6779	7447	6504	1279	1994	1091
R=3 × 100mg, S=200mg + 100mg, T=Test						

Table 4.35: Example 4.6: Williams Design for 4 Treatments

	Sequence ADBC							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
2	484	4190	509	4055	108.4	818.0	105.2	914.4
4	584	4134	450	3520	115.0	848.3	90.4	929.8
10	475	3596	350	2809	85.4	550.4	68.0	588.7
15	419	3430	454	3527	66.7	851.1	87.3	772.8
19	504	3635	429	4286	89.1	622.7	67.7	696.3
20	549	2727	314	3565	97.5	729.9	66.0	933.5
25	428	3174	389	3246	101.9	839.9	89.1	589.9
A=Reference Low, B=Test Low, C=Reference High, D=Test High								

Table 4.36: Example 4.6: Williams Design for 4 Treatments

	Sequence BACD							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
3	454	409	3571	3167	62.5	65.5	568.2	567.6
7	944	382	2830	2784	93.3	103.4	796.1	730.1
11	370	397	2399	1550	101.6	55.8	586.0	327.5
12	412	346	3010	2848	117.1	69.1	444.4	567.5
17	405	328	2574	2264	70.8	70.2	518.4	495.4
21	354	349	3249	2942	50.6	57.5	572.9	567.4
26	371	329	2427	2667	105.4	72.4	681.9	600.5
A=Reference Low, B=Test Low, C=Reference High, D=Test High								

Table 4.37: Example 4.6: Williams Design for 4 Treatments

	Sequence CBDA							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
6	3163	413	3069	345	689.1	94.6	652.1	58.2
8	3410	307	3009	370	554.6	61.5	675.1	87.2
9	3417	352	2975	376	686.6	56.8	606.0	59.8
14	3327	332	2826	350	629.1	86.0	718.8	87.1
18	2223	208	1759	232	563.2	67.7	584.1	74.0
22	2368	257	2104	274	540.2	50.5	464.1	59.2
28	3020	414	3022	419	652.7	59.3	607.2	79.1
A=Reference Low, B=Test Low, C=Reference High, D=Test High								

Table 4.38: Example 4.6: Williams Design for 4 Treatments

	Sequence DCAB							
	AUC Period				Cmax Period			
Sub	1	2	3	4	1	2	3	4
1	2942	2525	278	359	563.6	658.1	55.6	73.0
5	2740	2634	338	306	565.7	580.3	71.7	53.7
13	2897	2538	313	331	833.4	562.5	96.6	78.6
16	4513	4058	484	434	859.4	745.2	88.9	70.0
A=Reference Low, B=Test Low, C=Reference High, D=Test High								

Table 4.38: Example 4.6: Williams Design for 4 Treatments

	Sequence DCAB							
	AUC				Cmax			
	Period				Period			
Sub	1	2	3	4	1	2	3	4
23	2095	1987	233	199	388.1	471.6	54.2	25.2
24	3218	2705	365	367	635.6	643.3	66.3	62.1
27	2525	2672	238	316	471.2	557.0	29.4	51.4
A=Reference Low, B=Test Low, C=Reference High, D=Test High								

Table 4.39: Example 4.6: Tmax, Williams Design for 4 Treatments

	ADBC					BACD			
	Period					Period			
Sub	1	2	3	4	Sub	1	2	3	4
2	1.00	1.98	1.05	1.52	3	1.00	0.50	0.52	0.52
4	1.02	1.00	0.57	0.55	7	1.48	0.45	0.53	0.53
10	0.50	1.95	1.02	1.02	11	0.52	1.98	1.45	1.48
15	1.00	0.48	0.48	0.50	12	0.50	0.60	1.47	1.50
19	0.50	1.00	1.45	1.00	17	0.98	1.03	1.00	1.00
20	0.97	0.95	0.48	0.48	21	0.50	1.50	0.48	1.48
25	0.53	0.48	0.47	1.47	26	0.50	0.98	0.50	1.00
	CBDA					DCAB			
	Period					Period			
Sub	1	2	3	4	Sub	1	2	3	4
6	0.57	0.50	1.02	0.98	1	0.98	0.48	1.03	0.50
8	1.03	1.02	1.00	0.55	5	0.48	0.48	0.50	0.50
9	0.50	0.48	0.55	1.45	13	0.52	1.02	0.50	0.48
14	0.98	0.48	0.52	0.53	16	0.48	1.00	1.02	0.98
18	0.95	0.97	0.48	1.00	23	0.97	0.95	0.98	3.97
22	0.47	0.95	0.48	1.00	24	1.00	0.97	1.00	1.50
28	1.00	0.48	0.52	0.53	27	3.00	1.50	1.48	1.98
A=Reference Low, B=Test Low, C=Reference High, D=Test High									

CHAPTER 5

Dealing with Unexpected BE Challenges

or What one can do when some things that can go wrong, do go wrong....

In business, there is always a lot of talk about challenges and opportunities. These are really the same thing - a demanding task that requires a greater than normal commitment to see through to completion. In 'business-speak', when something goes wrong and has to be fixed or changed, this type of thing is typically viewed as a 'challenge'. A challenge is an opportunity that one does not want to work on as it has some sort of negative connotation associated with it. In contrast, an opportunity is a challenge that one does want to work on, as it has some sort of positive connotation. It is all a matter of one's perspective on the event in question. Either way, however, statistically speaking, it is probably going to require a lot of work.

When the FDA denies a claim of bioequivalence, there is a great deal of consternation at any given sponsoring company. Everyone usually knows such an event is coming, but it is like getting a big bill requiring immediate payment in the mail - if it came in tomorrow (or even better next week or next month), that would be preferable. My company was no exception, and our senior executives met quickly in one such instance to determine what to do. It was decided to repeat the study 'right away' (with some design enhancements). This was definitely referred to from the get-go as a 'challenge to the organization'.

On top of the dismay among the staff working on the project associated with not having our bioequivalence claim approved, this 'right away' action by senior executives generally represents an even greater challenge (and caused even more tangible consternation) among the staff who actually have to do this job. It is advisable not to tell a senior executive that something is impossible (if one values one's job), but often things like this can be very difficult if not impossible.

A human clinical trial of a drug product must have a written protocol (plan for the study) which must be unconditionally approved following review by an independent ethics review committee prior to any subject or patient being screened or dosed. If it involves a new chemical or bi-

ologic entity, this entity must be appropriately registered with the local regulatory agency of the appropriate government(s) if required; there is a lot of paperwork involved with these things. In addition, any human volunteer or patient involved in a trial must read and sign an informed consent before being enrolled in a trial and must be screened to ensure they are physically and mentally capable of taking part in the clinical trial. There are contracts with the site that must be reviewed by procurement and legal functions within the sponsoring company prior to signoff, and usually someone who is completely critical to at least one step in this process is on vacation, blissfully unaware of what is going on.

Thus, when something like the above study is to be done 'right away', that means the protocol, paperwork, and trial facilities for the study have to have been written and assembled and submitted several days or weeks ***ago*** *in order to get the study started as soon as possible. No one can reasonably be deemed a fortune teller and do such a thing in advance; hence, long hours and long days are the usual result of a challenge to try to have such a trial up and running 'Stat.'*

Senior executives are also the people responsible for resources at most companies and are often very surprised when they find out how long things really take. The good ones come down and lend a hand until the crisis is past, and we received a lot of senior-level attention on this occasion. Everything that could be done was done (in the space of three weeks), and the trial was good to go on the following Monday. However, we were all so busy working that we forgot to look out the window, an important omission on this occasion.

We finalized the protocol following ethics board review, completed all the regulatory paperwork, and finalized all the contracts. Then we all took a long deep breath and went home for the weekend to recover. Unfortunately, it rained all weekend as a hurricane was passing through the Caribbean and eastern USA.

We returned to work Monday morning to discover that despite our best efforts, the trial would not start as desired by our senior executive team. The hurricane had disrupted the shipping of supplies to the site, and we would have to reschedule. We had to delay the trial and resulting regulatory file dates. Some things just cannot be designed into or accounted for in models of bioequivalence trials.

Many other matters, however, can be controlled in design or modelled afterward to assess impact using technologies developed in the 20th century. In this chapter, we describe several such topics and methods for doing so. First, concepts behind failure to demonstrate bioequivalence are developed, followed by a brief discussion on the use of simulation for products failing to demonstrate bioequivalence. Next we turn to the use of restricted maximum likelihood models to explore the structure of variance *components and discuss techniques to assess the impact of carry-over in bioequivalence trials.*

5.1 Restricted Maximum Likelihood Modelling

The likelihood is the probability of observing the sample of data obtained in the trial, and is, given these data, a function of a set of specified parameters. For BE testing, the parameters of interest are the formulation, period, and sequences effects and any within- or between-subject variances. In trials where subjects get repeated exposure to a formulation, i.e., where the design includes sequences such as RTTR and TRRT, it is possible to estimate σ^2_{BT}, σ^2_{BR}, the between-subject variances of T and R, respectively, and the within-subject variances of T and R, σ^2_{WT}, σ^2_{WR}, respectively. The method of maximum likelihood (ML) determines the parameter estimates to be those values of the parameters that give the maximum of the likelihood. Restricted maximum likelihood estimation (REML) is a form of ML estimation that uses an iterative procedure where within each iteration there are two steps. A simplified description of REML is as follows. Using a first guess or estimate of the parameter values, the procedure keeps the values of the variance parameters fixed and estimates the formulation, period, and sequence effects. This is the first step. The residuals from this model are then calculated and used to reestimate the variance parameters. This is the second step. These steps are repeated until the values of the parameters do not change from one iteration to the next. The 'Restricted' in the name of the method arises because within each step, one set of parameters is fixed while the other set is estimated by maximizing the likelihood under the restriction imposed by the fixed set of parameters.

The usefulness of REML is that it can be used to estimate the between- and within-subject variances. The estimates, so obtained, are informative for the interpretation of the data, particularly when bioequivalence between T and R is not demonstrated. A second, and less important, property of REML is that it can be used when the data set is incomplete, i.e., when a complete set of logAUC or logCmax is not obtained from each subject. We illustrated such an analysis in Chapter 3. There, it will be recalled, the REML results were very similar to the results obtained from an analysis that used just those subjects that had a complete set of values. For more information on the properties of REML when the trial has a relatively small number of subjects, see [237] Chapter 6.

When a trial fails to show bioequivalence it is of interest to determine which factors (i.e., a difference in formulation means, unexpectedly high variability, or both) led to such a circumstance, and REML models may be used to explore data in such a context and in the presence of missing data. The use of REML models is also important in the context of individual and population bioequivalence, and we review the application of such methods in the next chapter.

REML has quite a long history [327], [181], [259], [47] and has been particulary useful for the analysis of repeated measurements [237], [307], [435], [250]. Readers interested in application in the bioequivalence setting should see [332], [335], and [457].

Obviously REML estimation cannot be done by hand. SAS code to perform these analyses is given in the following box.

For standard two, three, and four-period designs such as those found in Examples 3.1, 3.2, 4.5, and 4.6 (i.e., those where no formulation adminstration is replicated), analysis code may be found on the website in `exam1.sas - exam4.sas`, respectively. Some `proc mixed` code for Example 4.5 `exam3.sd2` is included here for illustration purposes:

```
proc mixed data=my.exam3
    method=reml ITDETAILS maxiter=200;
    class sequence subject period formula;
    model lnauct=sequence period formula
    /ddfm=KENWARDROGER;
    random subject(sequence);
    estimate 'T-R' formula -1 0 1
    /cl alpha=0.10;
    estimate 'T-S' formula 0 -1 1
    /cl alpha=0.10;
run;
```

Kenward and Roger's [246] denominator degrees of freedom are specified to ensure the correct degrees of freedom are used and that a good estimate of the standard error of $\hat{\mu}_T - \hat{\mu}_R$ is obtained.

Estimates relevant to ABE testing may be found in Table 5.1 for AUC, Cmax, and $T_{\frac{1}{2}}$ on the log scale. Note that Tmax was not analyzed using a log-transformation; thus differences expressed for Tmax are on the original scale in hours.

It will be recalled that while Example 3.1 demonstrated bioequivalence, Example 3.2 did not (see Chapter 3) due to reasons discussed later in this chapter. In Example 4.5, formulation T was not equivalent to R nor S with results indicative of a potentially bioinequivalent new formulation. In Example 4.6, bioequivalence was demonstrated at both high and low doses of drug product.

We have provided two additional data sets `exam5.sd2` and `exam6.sd2` on the website, that were obtained from trials that used the sequences (RTTR/TRRT) and (RTRT/TRTR), respectively. FDA-recommended code to analyse these [131] may be found in `exam5.sas` and `exam6.sas`, respectively. The proc mixed code for exam5 is given in the box below. The AUC and Cmax data for exam6 are given in Table 5.4, where we consider other interesting features of these data.

```
proc mixed data=my.exam5
    method=reml ITDETAILS maxiter=200;
    class sequence subject period formula;
    model lnauct=sequence period formula
    /ddfm=KENWARDROGER;
    random formula/type=FA0(2) subject=subject;
    repeated/group=formula subject=subject;
    estimate 'T-R' formula -1 1/CL ALPHA=0.1;
run;
```

Here, as before, the procedure mixed is called in SAS and estimates of the test and reference formulations differences are again computed using the estimate statement. Note, however, that different specifications are included for the random and repeated statements (cf., [131]). These are used as the replication of treatments within each subject permits the estimation of between- and within-subject variances for each formulation.

The random statement specifies that a particular choice for the variance structure should be assumed for σ^2_{BT} and σ^2_{BR} (Factor Analytic [368]) and that the variance associated with subject-by-formulation interaction,

$$\sigma^2_D = \sigma^2_{BT} + \sigma^2_{BR} - 2\rho\sigma_{BT}\sigma_{BR},$$

should be derived, where ρ is the between-subject correlation between the formulations. Estimates of these may be found in Table 5.2. Further discussion on σ^2_D and its use may be found in Chapter 6.

The repeated statement specifies that within-subject variance estimates should be derived for T and R formulations separately.

Average bioequivalence was demonstrated for exam5 with no evidence of a subject-by-formulation interaction ($\hat{\sigma}_D$ of 0.03 for AUC and 0 for C-max). Test and reference formulations were equivalent for AUC in exam6 with no evidence of a subject-by-formulation interaction. Note that for Cmax in data set exam6, however, in addition to a large increase in mean rate of exposure for the test formulation (0.4120, definitely indicative of bioinequivalence), there was evidence of a subject-by-formulation interaction as indicated by $\hat{\sigma}_D = 0.197$ (making it even more difficult to demonstrate bioequivalence as the confidence intervals will be wider by a factor directly proportional to this value). The data in exam6 had other aspects making it interesting statistically, and we will consider this data set in more detail later in the chapter.

Table 5.1 *REML Results from PROC MIXED for Standard Bioequivalence Designs*

Example	Data Set	Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% CI	$\hat{\sigma}^2_W$
3.1	exam2	logAUC(0-∞)	-0.0166	-0.0612, 0.0280	0.0110
		logAUC(0-t)	-0.0167	-0.0589, 0.0256	0.0099
		logCmax	-0.0269	-0.1102, 0.0563	0.0384
		logThalf	0.0108	-0.0630, 0.0846	0.0301
		Tmax	0.0008	-0.2191, 0.2208	0.2676
3.2	exam1	logAUC(0-∞)	0.0940	-0.0678, 0.2482	0.1991
		logCmax	0.0468	-0.0907, 0.1843	0.1580
4.5	exam3	logAUC(0-t)	$T - R$=0.1497	0.0859, 0.2136	0.0440
			$T - S$=-0.1912	-0.2554, -0.1270	
		logCmax	$T - R$=0.2597	0.1726, 0.3468	0.0846
			$T - S$=-0.2088	-0.2964, -0.1212	
		Tmax	$T - R$=-0.5060	-0.7223, -0.2897	0.5220
			$T - S$=-0.0713	-0.2888, 0.1462	
4.6	exam4	logAUC(0-∞)	$B - A$=0.0047	-0.0545, 0.0638	0.0177
			$D - C$=-0.0362	-0.0953, 0.0230	
		logCmax	$B - A$=-0.0355	-0.1171, 0.0461	0.0336
			$D - C$=-0.0301	-0.1117, 0.0515	
		logThalf	$B - A$=0.0590	0.0019, 0.1160	0.0164
			$D - C$=0.0258	-0.0313, 0.0828	
		Tmax	$B - A$=0.0243	-0.1855, 0.2341	0.2223
			$D - C$=0.1057	-0.1041, 0.3155	
B, D, T = Test Formulations A, C, R, S = Reference Formulations					

5.2 Failing BE and the DER Assessment

For some drug products, even if one tries time and time again to demonstrate bioequivalence, it may be that it just cannot be done. A common misconception is that this means the test and reference formulations are 'bioinequivalent,' in that they deliver different pharmacokinetic profiles causing different pharmacodynamic response. This is not necessarily the case.

An example of a potentially bioinequivalent test product is presented in Figure 5.1. The important thing to note is that the measure of centrality in addition to the bulk of the distribution falls outside the average bioequivalence confidence limit for logAUC. Implicitly, for a product to demonstrate bioequivalence, its true measure of centrality must fall within the limits. Otherwise, it will be next to impossible (or a Type 1 error) for such a product to demonstrate bioequivalence. In this case,

Table 5.2 *REML Results from PROC MIXED for Chapter 5 Examples of Replicate Designs*

Data Set	Endpoint	$\hat{\mu}_T - \hat{\mu}_R$	90% CI	$\hat{\sigma}_D$	$\hat{\sigma}^2_{Wi}$
exam5	logAUC(0-∞)	0.0357	0.0096, 0.0617	0.026	R = 0.0065 T = 0.0117
RTTR/TRRT	logCmax	-0.0918	-0.1766, -0.0070	0	R = 0.0404 T = 0.0587
exam6	logAUC(0-t)	0.1004	0.0294, 0.1714	0	R = 0.1202 T = 0.0756
RTRT/TRTR	logAUC(0-t')	0.1043	0.0326, 0.1760	0	R = 0.1212 T = 0.0818
	logCmax	0.4120	0.2895, 0.5346	0.197	R = 0.3063 T = 0.2689
	Tmax	-0.1716	-0.5765, 0.2334	0	R = 0.9417 T = 3.8484

T = Test Formulations
R = Reference Formulation

the estimated $\mu_T - \mu_R$ tells us that it is most unlikely we will ever be able to demonstrate bioequivalence.

Contrast this with the fitted normal densities of Example 3.2 in Chapter 3. Here the measure of centrality lies within the average bioequivalence acceptance limits, but slightly too much of the distribution lies outside to conclude the test and reference formulations are bioequivalent. These formulations are not bioinequivalent, but insufficient evidence has been provided to show that they are.

Formulations may fail to show bioequivalence for several reasons:

1. The estimated $\mu_T - \mu_R$ lies too far from zero,
2. Variation is greater than expected, resulting in too wide a confidence interval for $\mu_T - \mu_R$,
3. Insufficient sample size is used (also yielding too wide a confidence interval for $\mu_T - \mu_R$),
4. Or some combination of these.

In Example 3.2 of Chapter 3, all three factors combine to contribute to the failure to demonstrate bioequivalence. The difference in formulation means, $\hat{\mu}_T - \hat{\mu}_R$, was estimated to be approximately 0.1 for logAUC (on the natural scale, 1.1) while the study had been designed under the assumption that $\mu_T - \mu_R$ would be no greater than ± 0.05. Also, $\hat{\sigma}_W$ was estimated to be approximately 0.45 while the sample size had been chosen in expectation of a σ_W of 0.3. The combination of these two factors in combination with the a priori choice of sample size resulted

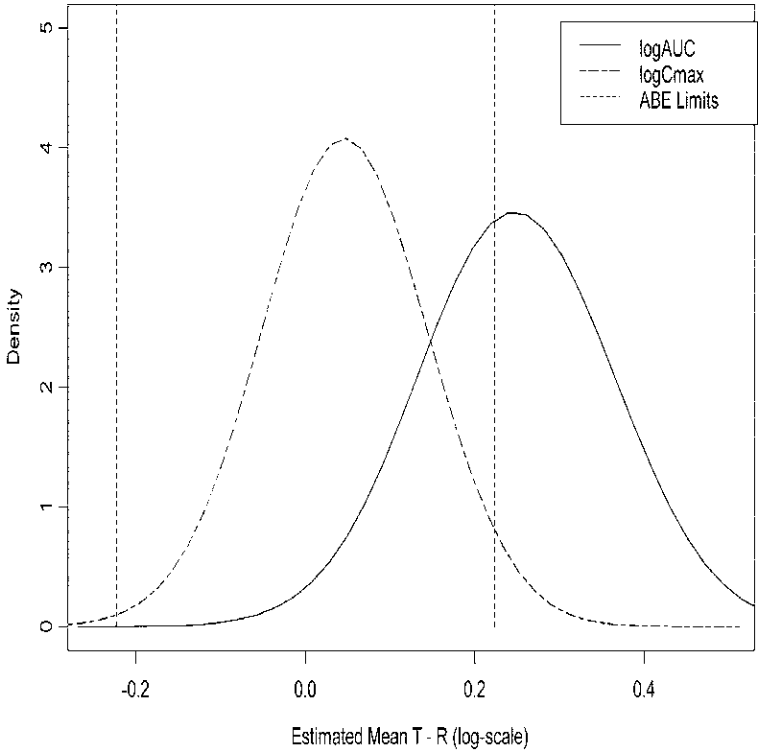

Figure 5.1 *A Potentially Bioinequivalent Test and Reference Product: Fitted Normal Densities for* $\hat{\mu}_T - \hat{\mu}_R$

in a failure to demonstrate bioequivalence. However, given the observed magnitude of these factors a better designed follow-up study might be able to show bioequivalence successfully. Some might refer to the study as having been 'underpowered' implying that insufficient sample size was utilised; however, all three factors contributed to the failure to demonstrate bioequivalence.

Insufficient sample size can result in confidence intervals that are wide in bioequivalence trials, making it difficult to demonstrate bioequivalence. Note, however, that in such failed trials the confidence interval is quite informative [187]. In the case of a failed bioequivalence trial, the confidence interval may be regarded as expressing a plausible range of values for the true $\mu_T - \mu_R$. In the case of Example 3.2, the confidence interval for AUC (recall this is $exp(\hat{\mu}_T - \hat{\mu}_R)$) was (0.94-1.29) with the probability of any given value of $exp(\mu_T - \mu_R)$ decreasing as it becomes further away from 1.1. It is possible therefore that if we repeated the trial using the same design and sample size that we would observe $exp(\hat{\mu}_T - \hat{\mu}_R)$ of as low as 0.94 and as high as 1.29! Indeed in a previous relative bioavailability study (of similar design but lower sample size) an estimate of 0.95 for $exp(\hat{\mu}_T - \hat{\mu}_R)$ had been observed.

Note that in Example 3.2, as a 2×2 cross-over design was used, hence $\sigma_{WT} = \sigma_{WR} = \sigma_W$. These variances are confounded in this design, and we can neither test nor estimate whether $\sigma_{WT} = \sigma_{WR} = \sigma_W$. If a replicate design had been used, it would be possible to separately model the magnitude of intra-subject and inter-subject variation for each formulation. Such a design and analysis might be desirable if we suspected, for instance, that the new formulation resulted in more intra-subject variation than the reference formulation.

Failure to demonstrate bioequivalence is therefore different but related to bioinequivalence. Only in cases where sample size is very large and point estimates for δ lie outside the acceptance bounds would one definitely conclude bioinequivalence was observed. Bioinequivalence is thus quite rare, but failure to demonstrate bioequivalence can occur quite often. In the latter case, it is generally possible to repeat the study or use a more powerful design to attempt to successfully demonstrate bioequivalence.

Turning now to the implications of failure to demonstrate bioequivalence, successful demonstration is not always necessary in regulatory science to secure approval of a new product. For certain new agents, rate and extent of exposure can change in a new formulation relative to that used in clinical trials. For a new product's first regulatory application (i.e., a product invented by the sponsor representing a new chemical or biological entity's first New Drug Application at the FDA, for example), a drug might not need to clearly demonstrate bioequivalence to the full regulatory standard. The FDA's guidance on this follows:

> Where the test product generates plasma levels that are substantially above those of the reference product, the regulatory concern is not therapeutic failure, but the adequacy of the safety database from the test product. Where the test product has levels that are substantially below those of the reference product, the regulatory concern becomes therapeutic efficacy. When the variability of the test product rises, the regulatory concern relates to both safety and efficacy, because it may suggest that the test product does not perform as well as the reference product, and the test product may be too variable to be clinically useful.
>
> Proper mapping of individual dose-response or concentration-response curves is useful in situations where the drug product has plasma levels that are either higher or lower than the reference product and are outside usual BE limits. In the absence of individual data, population dose-response or concentration-response data acquired over a range of doses, including doses

> above the recommended therapeutic doses, may be sufficient to demonstrate that the increase in plasma levels would not be accompanied by additional risk. Similarly, population dose- or concentration-response relationships observed over a lower range of doses, including doses below the recommended therapeutic doses, may be able to demonstrate that reduced levels of the test product compared to the reference product are associated with adequate efficacy. In either event, the burden is on the sponsor to demonstrate the adequacy of the clinical trial dose-response or concentration-response data to provide evidence of therapeutic equivalence. In the absence of this evidence, failure to document BE may suggest the product should be reformulated, the method of manufacture for the test product be changed, and/or the BE study be repeated. [135]

If bioequivalence has not been demonstrated for a new product, the task then is to model exposure's (AUC, Cmax) relationship to efficacy and safety in patients using the reference formulation's clinical data. If therapeutic equivalence can be shown for such an exercise, then approval may be obtained. In the knowledge of the extent to which the test formulation changes exposure (measured in a bioequivalence study or studies), one may *simulate* what a change of the magnitude observed for AUC and Cmax for the test formulation would be produced in terms of patient response in clinical use.

We refer to this type of modelling and simulation procedure as the DER (Dose-Exposure-Response) assessment. Modelling of bioequivalence data was covered in Chapters 3 and 4. We will develop the basic ideas behind simulation in the next section, and modelling for the DER assessment will be developed in detail in Chapters 7 through 10.

This simulation-based procedure provides regulators with a technique to assess whether the issue in manufacturing poses a risk to the patients using the new product. Note, however, that the DER assessment is limited in scope of application ***to only new (i.e., innovative) products***. Existing marketed products may not apply such a procedure and must demonstrate average bioequivalence to have access to market (in most cases). There are always exceptions to such a rule, but such exceptions are very rare.

Bear in mind that regulators use average bioequivalence testing as a tool for measuring manufacturing quality. It is not the only tool which may be applied (see Chapter 2), and the extent of its rigor in its application is dependent on how many people are using the product in the marketplace.

For a new innovator product, relatively few numbers of patients (only those volunteering for clinical trials, see Chapter 2) will have received the drug. However, when a drug is allowed marketplace access by regulators, the number of patients exposed to drug increases exponentially. Small changes in the PK for a new innovator product may not result in increased risk to patients in the marketplace using the drug for the first time, and this is studied using the DER assessment. Regulators therefore are free to use their informed judgment in permitting market access for these innovative products.

When a manufacturer makes changes to a marketed formulation or multiple companies begin to market new formulations (at the innovator's patent expiration), there is little room for such judgment. Many people are presumed to be at risk, and the regulators must ensure that when the patients use the new formulations their safety and efficacy are protected. When millions of people are using a drug, even a very small change in exposure for a small percentage of patients may result in many people being placed at risk.

Conservative application of the average bioequivalence standard is therefore the rule once an approved drug is on the market, and regulators have little to no freedom to change bioequivalence limits. With few exceptions [17], rigorous application of the 0.80-1.25 limits has protected public health and individual patients using new formulations.

The rationale for this regulatory conservatism is well documented. Hauck et al. [197] showed that allowing wider than the usual acceptance limits (0.80-1.25) allowed larger changes in rate of exposure. This change could result in a less acceptable safety profile for a new formulation (i.e., more undesirable side effects) than the reference formulation. Anderson and Hauck [10] showed that rigorous application of the ABE acceptance criteria protects public health when multiple new formulations enter the marketplace at patent expiration.

Application of the DER assessment is therefore limited in scope to innovator products entering the market for the first time. We now turn to the topic of simulation in order to develop how one goes about a DER assessment.

5.3 Simulation

We introduce simulation here to develop the concepts behind its use and application in clinical pharmacology research. In practice in bioequivalence trials, it is not often needed. Most modern companies have manufacturing well under control by the time of a regulatory application. Bioinequivalence is quite infrequent, and failed bioequivalence studies are becoming rare with the advent of customization and automation in drug development manufacturing.

Simulation is simply defined as 'a means of creating data using the computer without going to the trouble of actually doing a study and collecting observations.'

This approach assumes we know the truth about the parameters in which we are interested. In bioequivalence, for example, we assume we know the true values of μ_T, μ_R, σ_B^2, σ_W^2, and the magnitude of any period and sequence effects. A random data generator can then be used in SAS, for example, to simulate PK data.

For example purposes, assume a bioequivalence study has failed, showing a 10% decrease in the new formulation AUC relative to the reference formulation. A logAUC of 175 is needed for the product to be efficacious in killing bacteria, and concern might exist that if a patient were switched from the reference to the new formulation in the course of a trial that the product might fail to demonstrate efficacy. The reference product had an average logAUC of 200 ($\mu_R = 200$), and the new formulation was observed to have an average logAUC of 180 ($\mu_T = 180$). We know from previous experience that between-subject variance for logAUC is 0.18 with a within-subject variance of 0.09.

SAS code (see the Technical Appendix) can then be used to generate simulated PK data for a cross-over study. Here we set $\mu_T = \ln 180$, $\mu_R = \ln 200$, $\sigma_B^2 = 0.18$, $\sigma_W^2 = 0.09$, and set period effects to null. Sequence effects are also set to null. LogAUC data for 2500 simulated subjects switched from the reference to the new product and vice versa in a 2×2 cross-over are output in the SAS data set `simulate`. This data may then be modelled using REML or the approaches of Chapter 3 to assess the statistical properties of such data.

In statistics, such simulations are used often in working practice. Simulated data are generated, and plugged into various methods of analysis under consideration to assess the properties of the statistics being considered. Statisticians may use such techniques to assess the degree of bias (the degree to which $\hat{\delta} \neq \delta$ for example) and precision ($(\hat{\delta} - \delta)^2$ for example). Statisticians also use such techniques to evaluate 'what if' scenarios. For example, the presence of two or three subjects with very unusual data points may easily be included in a simulation to assess their impact on the probability of demonstrating average bioequivalence.

Other branches of clinical pharmacology use simulation for other purposes - e.g., the DER assessment described in the last section. Response data are collected and used to develop models to relate exposure to response (see Chapter 9). In our example, we would evaluate the number of subjects achieving an efficacious response (logAUC$>$ 175) on the reference formulation and of these subjects assess how many subsequently showed an efficacious response on the test formulation. Note that these findings might also lead one to wish to increase the dose! However, one might be constrained in that a logAUC greater than 300 (for example) might be associated with an undesirable side-effect. Consideration of such is left to the reader.

This is a simplistic modelling and simulation example, but the concepts may be applied in more complex situations. We will develop the concepts supporting such tools in Chapters 7-10.

Clinical pharmacologists work with statisticians to develop such models and use simulations to predict what might be observed in future trials. This is a powerful tool; however, we need to have a care to monitor the assumptions being made in such an exercise. Results are highly dependent on the chosen model parameters, and life has a way of being more complex than any simple model can hope to describe.

5.4 Data-based Simulation

The first data exploration technique we will consider is a technique some attribute to R.A. Fisher [176] and developed in great detail in an excellent book by Efron and Tibshirani [96]. We encourage readers interested in application of this technique to explore these and other books [394] and publications (e.g., [395]-[396]) on the topic.

In this section, we will dwell on the application of the **bootstrap** in bioequivalence. The reader will note its utility as a general data exploration tool, and it will become very handy in our exploration of other clinical pharmacology data in Chapters 7-10.

The bootstrap is 'a computer based method for assigning measures of accuracy to statistical estimates' [96]. Essentially, we recognize that the sample of data from our trial is *only* a sample from a far larger population (which we obviously cannot sample exhaustively - it is too big, see Chapter 1). The data from each subject is simply a sample of what we would see if we studied that subject again and again. We could even drill down further and look at each individual period's results for each subject as a sample of what we would see if we repeated each period within each subject again and again. However, for this section, we will choose to apply the bootstrap at the subject level, maintaining the actual number of subjects observed within each sequence in accordance with recent draft guidance on the topic [122].

Bootstrapping is accomplished by randomly sampling, with replication, from the original data set of n subjects. One picks a subject at random from the data set, includes that data in the analysis data set, replaces the subject, picks again, replaces the subject, etc., until one has a new data set with n subjects. The same subject may appear in the bootstrap data set more than once.

One does this a large number of times to accumulate a set of r bootstrap data sets. The number r is arbitrary but should be pretty large, in general, at least $r \geq 1000$. The chosen method of analysis is then applied to each of the r bootstrap data sets, and a record of each of the r fitted sets of parameters is kept. For any given parameter, the r sets of estimates may be used to estimate moments of the parameter of interest such as its mean and variance.

Obviously, one cannot bootstrap this set of r data sets by hand, and application of this technique was constrained until modern computing power became available in the 1980s-1990s. Some modern software packages (e.g., SPLUS) offer automated bootstrapping routines, and bootstrapping is easily accomplished in SAS via use of the `MACRO` SAS language.

A SAS macro used for this purpose may be found on the website accompanying this book. Bootstrap samples are generated calling the SAS macro `bootstrp` from a chosen data set. Note that a `seed` value is input. This is a random number chosen to tell SAS where to begin sampling and allows one to reproduce the results if the program needs to be rerun. If a `seed` is not provided, SAS uses the clock to automatically determine where to start. The `bootstrp` macro then samples from the data set in the manner described above and outputs data sets $r = 1$ to `nrep` where `nrep` is the number of r bootstrap data sets desired. The number chosen in the examples is r=`nrep`=2000.

One then derives the statistic of interest for each bootstrap data set. For `exam1 - 6` we will estimate the 90% confidence interval for $\mu_T - \mu_R$ for the purposes of providing an example, though any statistic may be treated in this manner. For this exercise, we will be interested in estimating the proportion of cases among the bootstraps where a conclusion of BE may be made.

Following some data manipulations, the output bootstrap data sets are then each used to estimate a 90% confidence interval for $\mu_T - \mu_R$ using `proc mixed` as shown in Chapter 3. If the confidence interval falls within $\mp \ln 1.25$ for both $\ln AUC$ and $\ln Cmax$ then an overall 'success' is registered for that bootstrap data set.

We can see that it is unlikely a repeat of study in Example 3.2 (`exam1.sd2`) would be successful. See Table 5.3. Overall only 38% of bootstrapped data sets resulted in a conclusion of bioequivalence. While the percentage of Cmax data sets being bioequivalent was relatively high (at 67%), only 41% of bootstrapped data sets were successful for AUC.

One could also use this tool to evaluate 'what if' scenarios - e.g., what if we changed the sample size to $n = 20$ subjects? One could also run the bootstrap procedure repeatedly to obtain a confidence interval for the odds of a successful repeat of a bioequivalence trial. This sort of exercise is left to the reader.

As a caution, we advise that when using complex models like those currently employed in a bioequivalence testing, users of the bootstrap should take care to ensure that their findings are robust to the incidence of non-convergence in the bootstrapped data sets. Note that the REML

Table 5.3 *Number of Successful BE Trials*

Data Set	Comparison	AUC	Cmax	Overall
`exam1.sd2`	T-R	41%	67%	38%
`exam2.sd2`	T-R	100%	99%	99%
`exam3.sd2`	T-R	61%	0%	0%
`exam3.sd2`	T-S	16%	5%	4%
`exam3.sd2`	S-R	0%	0%	0%
`exam4.sd2`	B-A	100%	98%	98%
`exam4.sd2`	D-C	100%	100%	100%
`exam6.sd2`	T-R	88%	0%	0%

model used to examine `exam5.sd2` failed to converge in SAS when bootstrapped on a very large number of occasions due to the issues involving the magnitude of variances described in [332]. Therefore, results should be interpreted with caution and are not presented in Table 5.3. The analysis for `exam6.sd2` also failed to converge on a very limited number of occasions (less than 4% of the bootstraps for AUC, and less than 1% for Cmax). Modification to SAS code (see bootstrap_exam1 - 4.sas) may be necessary to ensure enough computer memory is available to run the model repeatedly or to ensure the model converges adequately.

We note that, while the bootstrap is a nice, easy to implement data exploration tool given modern computing power, it is important to note that the bootstrap sampling introduces randomness in to the results This randomness has implications. In some cases [335], coverage probability of confidence intervals generated using the bootstrap may be lower than expected, leading to an increased possibility of a Type 1 error (see Chapter 1). Therefore, while it is a useful tool for exploring data, caution should be applied when using any findings for making claims in regulatory submissions. Those doing so should be prepared to ensure regulators that the risk of a Type 1 error is maintained at a level acceptable to their public's health.

Although not utilised here, we further note that the bootstrap is a very powerful tool for model validation [179].

5.5 Carry-over

When carry-over is mentioned in bioequivalence studies, it refers to the occurrence of a nonzero plasma concentration of drug in a sample prior to dosing. As such it complicates the analysis of bioequivalence data, by aliasing or biasing the assessment of changes between formulations. To prevent this, a washout period (of at least five half-lives) is employed to prevent such occurrences.

Carry-over is very unusual but not unknown in bioequivalence studies and can arise from a variety of factors. Some are:

1. Long-half life drugs (with inadequate, too short, washout duration),
2. Serendipitous inclusion in the trial of subjects who poorly metabolize or eliminate the drug,
3. Random occurrences (possibly due to assay problems).

Statistical tests are available to test for carry-over and to evaluate its impact on these changes in formulation means [237]; however, in keeping with comments made in Chapter 3, and previous findings [388], we do not recommend that those analysing data from bioequivalence studies carry out statistical tests for the presence of carry-over [237], [389]-[390]. We will therefore confine discussion to practical issues and analyses that may be considered when pre-dose concentrations are detected. This would signal that carry-over was present in the bioequivalence design, and we assume that statistical tests will not be used to assess its impact in keeping with [389]-[390] and [237].

As a practical matter, even if a more than adequate washout is used, there will be instances where pre-dose concentrations in periods after the first are non-null. The example to be considered was a drug that had been on the market for so long that its development predated pharmacokinetic assessment! The plant where the formulation had been manufactured (for many, many years) was closing, and the machinery that made the drug was packed and shipped to another site to continue manufacture, and the people who ran the machines at the old site retired. Therefore, the job was to prove that manufacturing at the new site with the new people but old equipment was to the same quality as the old (closed) site by use of a bioequivalence test.

In designing this bioequivalence study, the complete lack of pharmacokinetic data was problematic on this occasion as we had no basis on which to decide the length of a sufficient washout period, and there was insufficient time to run a pilot study In the end, the study was designed based on an educated guess about what washout was needed from pharmacodynamic action of the product, but it turned out our guess undershot the needed duration. The example represents a worst-case scenario in that 48 pharmacokinetic profiles (for 27 subjects of 54 participating) were identified as having pre-dose concentrations in excess of the pharmacokinetic assay's lower limit of detection. AUC and Cmax data are listed below for this replicate design. AUC and Cmax values marked with a 'C' denote those where a pre-dose plasma concentration was non-zero and in excess of the pharmacokinetic assay's lower limit of quantification.

Table 5.4: Example 5.1: AUC and Cmax Data from a Replicate Cross-over Study Design with Test and Reference Formulations and Carry-over (C)

Subject	Seq	Period 1	Period 2	Period 3	Period 4
AUC					
1	RTRT	812.6	1173.7C	889.1	620.1
2	TRTR	216.3	338	502.8C	398.6
3	RTRT	545.1	542.9C	.	.
4	TRTR	632.6	520C	716.7C	860.4C
5	RTRT	400	223.8C	173.7	289.7C
6	RTRT	102.1	185.3	42	88.3
7	TRTR	596	659.3	543.8	662.9
8	TRTR	402.4	359.8	590.8	444.3
9	TRTR	456.7	378.4	477.5	407.9C
10	RTRT	304.5	351.5C	520.2C	335.7C
11	TRTR	500.7	323C	416.3C	525.1C
12	RTRT	176.1	710.7	409.5	645.5
13	TRTR	160.6	218	170.1	124.6
15	RTRT	562.4	490.4C	504.7	675.9
16	TRTR	756	606.8	477.4	626.8
17	RTRT	207.5	271.6	173.7	240.5
18	RTRT	571.3	705.2	619	633.6
19	TRTR	511.9	549.7	388.2	141
20	TRTR	124	91.9	113.3	59.5
21	RTRT	536.1	595.2	445.5	521.5C
22	TRTR	239.7	265.1C	445.9	433.2
23	TRTR	609.6	371.6C	511.3	432.7C
24	RTRT	449.9	860.4C	606.8	577.2C
25	RTRT	192.5	220.1	233.1	227
26	TRTR	764.4	508.8	757.8	449.4
27	TRTR	151.9	194.8	.	.
28	RTRT	568.1	321.1	338.3	403.6C
29	RTRT	735.6	634.5C	1244.2C	641.9
30	TRTR	429.1	391.8C	316.9	335.1C
31	RTRT	307.4	481.8	346.6C	369.7C
32	TRTR	409	514.6C	763.1C	406.5C
33	TRTR	271	221	296.5	463.7
R=Reference, T=Test C=Carry-over Concentration at Baseline					

Table 5.4: Example 5.1: AUC and Cmax Data from a Replicate Cross-over Study Design with Test and Reference Formulations and Carry-over (C)

Subject	Seq	Period 1	Period 2	Period 3	Period 4
34	RTRT	292.9	431C	448.5	267.8C
35	RTRT	217.2	332.2	103	127.5
36	TRTR	290.8	208.6	243.7	489.8
37	TRTR	297.2	502	320.4	334.3
38	TRTR	163.8	232.1	636.9	434.9
39	RTRT	368.3	292.6C	446.1	222.3C
40	RTRT	193.7	202.8	255.2	244.3
42	TRTR	534.1	243.1	418.4	441.9
43	TRTR	355.1	415.2	382.7	334
44	RTRT	102	282.5C	245.6	286.2C
45	TRTR	320.5	233.9	331.7	260.5
46	RTRT	223.6	645.4	349	507.4C
47	TRTR	504.5	289.9	550.7C	244.2
48	RTRT	615.8	732.1C	620.9	665.2C
49	RTRT	898.4	924.9C	398.3	828.3C
50	RTRT	410.4	329.2	449.4	442.1
52	TRTR	237	505C	496.3	580.6C
53	RTRT	332.4	273.6	525.3	293.3
54	RTRT	185.2	222.9	182.1	194.1
55	TRTR	246.9	620.9C	678.3	752.2C
56	TRTR	235.4	190.4	318.3	248.4
57	RTRT	180.6	174.7	102.9	117
Cmax					
1	RTRT	99.85	204.09C	170.94	112.78
2	TRTR	29.06	50.48	35.15C	55.71
3	RTRT	67.69	41.73C	.	.
4	TRTR	91.25	43.86C	168.78C	61.04C
5	RTRT	40.05	25.17C	24.48	86.49C
6	RTRT	28.76	24.83	9.27	10.89
7	TRTR	257.1	79.04	127.92	81.8
8	TRTR	136.27	158.86	148.97	82.41
9	TRTR	65.48	87.84	64.57	58.01C
10	RTRT	34.35	52.26C	142.92C	58.48C
11	TRTR	31.49	37.07C	80.9C	33.62C
12	RTRT	18.94	161.34	118.89	246.57
13	TRTR	29.61	43.15	27.71	13.11

R=Reference, T=Test
C=Carry-over Concentration at Baseline

Table 5.4: Example 5.1: AUC and Cmax Data from a Replicate Cross-over Study Design with Test and Reference Formulations and Carry-over (C)

Subject	Seq	Period 1	Period 2	Period 3	Period 4
15	RTRT	28.35	98.5C	78.22	140.54
16	TRTR	168.76	174.94	117.31	52.18
17	RTRT	19.18	94.92	21.39	65.45
18	RTRT	66.63	134.69	78.1	78.51
19	TRTR	32.23	70.06	32.15	43.11
20	TRTR	9.34	11.74	49.23	18.42
21	RTRT	42.11	37.82	39.87	116.79C
22	TRTR	38.02	16.79C	38.58	83.82
23	TRTR	199.07	52.14C	118.47	72.04C
24	RTRT	32.53	276.86C	118.65	156.33C
25	RTRT	21.96	38.97	22.26	54.16
26	TRTR	74.24	35.76	39.27	36.28
27	TRTR	19	20.61	.	.
28	RTRT	110.87	55.64	50.06	84.6C
29	RTRT	50.08	58.79C	181.53C	144.26
30	TRTR	31.85	74.88C	54.88	19.18C
31	RTRT	87.21	88.75	90.07C	132.92C
32	TRTR	30.86	70.84C	208.2C	65.25C
33	TRTR	86.01	41.85	67.86	79.81
34	RTRT	18.07	33.37C	21.48	20.87C
35	RTRT	18.69	174.55	17.06	32.01
36	TRTR	38.27	40.31	31.56	20.64
37	TRTR	49.81	66.64	17.8	25.94
38	TRTR	34.56	16.37	114.3	29.58
39	RTRT	52.59	57.88C	48.58	47.24C
40	RTRT	29.3	78.33	21.72	49.27
42	TRTR	136	33.75	104.12	35.03
43	TRTR	64.55	34.04	52.37	41.67
44	RTRT	22.14	63.5C	9.38	16.3C
45	TRTR	26.35	37.2	76.26	24.6
46	RTRT	27.02	167.28	20.35	121.92C
47	TRTR	118.91	49.27	166.61C	35.86
48	RTRT	60.94	100.47C	26.17	98.08C
49	RTRT	164.01	180.01C	25.21	97.02C
50	RTRT	59.7	43.65	102.47	40
52	TRTR	30.55	63.9C	39.17	40.75C

R=Reference, T=Test
C=Carry-over Concentration at Baseline

Table 5.4: Example 5.1: AUC and Cmax Data from a Replicate Cross-over Study Design with Test and Reference Formulations and Carry-over (C)

Subject	Seq	Period 1	Period 2	Period 3	Period 4
53	RTRT	39.96	56.47	42.11	38.75
54	RTRT	18.34	16.09	21.5	9.57
55	TRTR	42.2	106.69C	150.52	115.15C
56	TRTR	39.15	13.79	122.03	62.32
57	RTRT	9.1	58.44	12.74	18.33
R=Reference, T=Test C=Carry-over Concentration at Baseline					

The average contribution of the nonzero pre-dose concentrations in these subjects relative to the magnitude of Cmax observed in that period was only 1.5%. Thus, for the majority of subjects the presence of carry-over could be deemed negligible as a practical matter. However, two subjects had concentrations of approximately 5.25% and 5.05% relative to the Cmax observed in that period. For the purpose of this example, Subject 1 (Period 2, pre-dose concentration of 10.7) and Subject 10 (Period 3, pre-dose concentration of 7.2) will be deemed to have had such observed data. Results of such magnitude would lead one to consider whether this could have an impact on inference.

One regulatory guidance is quite clear on how to handle such data. The FDA guidance [135] recommends that:

> If the pre-dose concentration is less than or equal to 5 percent of Cmax value in that subject, the subject's data without any adjustments can be included in all pharmacokinetic measurements and calculations. If the predose value is greater than 5 percent of Cmax, the subject should be dropped from all BE study evaluations. [135]

This guidance [135] tacitly assumes that the occurrence of carry-over of sufficient magnitude to impact inference is random in line with recent publications on the topic [389]-[390], [237]. Indeed, even in the worst-case scenario of carry-over described above, only two subjects had concentrations of such magnitude, and exclusion of subjects 1 and 10 does not materially effect statistical inference. Confirmation of this finding is left to the reader using the SAS code introduced earlier.

This approach has the benefit of simplicity, and given the sparsity of the occurrence of relevant carry-over, it is expected that its application will suit most circumstances. We will dwell on an alternative approach for (rare) situations when such might not be suitable. Additionally, the handling of such data by non-FDA regulatory bodies, for example, is not standardized in guidance, and other agencies might request other approaches to analysis.

We will consider adjustment of data for pre-dose concentrations where the magnitude is of sufficient magnitude to warrant concern. Consider subjects 1 and 10 in the example. As the pre-dose concentration has been assayed, it is reasonable to presume that the effect on subsequent observed plasma concentrations is additive relative to the dose received in the period under study as drug on board pre-dose is simply in the process of being eliminated from the body. One could 'slice' a portion of the concentration data from the pharmacokinetic concentration versus time profile and estimate an adjusted Cmax and AUC for use in analysis. For Example 5.1, these data are presented in Table 5.5.

Table 5.5 *Example 5.1: Adjusted Cmax and AUC Data for Subjects 1 and 10*

Subject	Period	Adj. AUC(0-t)	Adj. Cmax
1	2	1045.3	193.39
10	3	433.8	135.72

These adjusted data were derived by subtracting the pre-dose concentration from Cmax and by removing an estimated area from the AUC. For the purposes of this example, the t in AUC(0-t) was defined as occurring at 24 hours post dose, and the area 'sliced' from the full AUC was derived according to the equations of Chapter 1 consistent with a half-life of 24 hours. In more complex pharmacokinetic profiles, other calculations might be more appropriate. The changes in the data introduced by 'slicing' adjustment did not impact statistical inference (confirmation of this is left to the reader) relative to statistical analysis including their original data.

Data manipulation in such a manner may have two unintended effects on bioequivalence testing. It certainly introduces more variation into our model, as adjustment using 'slicing' does not take into account the error implicit to pharmacokinetic measurement (due to assay and within-subject variability).

Additionally, data manipulation may introduce bias as only some data for subjects in certain periods were adjusted. Note in the example above, subjects with very low pre-dose concentrations were present but were neglected as their values were minimal (around 1.5% of Cmax), and only data for certain periods were adjusted (see Subject 10). As average bioequivalence is a within-subject assessment (each subject serves as their own control), if a subject experiences carry-over sufficient to warrant adjustment in one session, it is more appropriate to adjust all sessions for any relevant pre-dose concentrations for a given subject regardless of magnitude. In the case of the above example, one should consider adjusting data for Subject 10 in Periods 2 and 4 even though pre-dose concentrations in those sessions were only approximately 1.5% of Cmax. Alternatively, if the FDA 'data-reduction' approach [135] is used, all of the subject pharmacokinetic data should be excluded from analysis to avoid the introduction of bias or increased variation.

Those using such adjustment techniques should be explicit about what adjustments were made and the process followed. Regulatory acceptance of such a procedure is unknown and falls outside the scope of current international guidance (other than the FDA's, cf., [135], where an alternative procedure described above is recommended).

Note also that while not recommended in bioequivalence testing, statistical models for the assessment of differential carry-over in bioequivalence testing in cross-over designs [237] may detect such 'slicing' of the data as a factor consistent with the statistical detection of carry-over from such data. Therefore, further care is recommended if such models and data manipulation are applied.

5.6 Optional Designs

The most common study design applied to bioequivalence testing is the 2×2 design, already described in great detail. In cases where one dosing regimen is to be marketed relative to the multiple formulations used in confirmatory trials (e.g., a 300 mg dose is to be marketed but must be confirmed as bioequivalent to a 200 mg tablet with a 100 mg tablet and to three 100 mg tablets), it may be necessary to extend this design to consider more than two regimens in a BE trial. Other trials might include four periods if bioequivalence was to be evaluated between two formulations at a low and at a higher dose, for example. Such is required by certain nations [58] when dose-proportionality (see Chapter 7) is not demonstrated.

Such designs are simply an extension of the 2×2 design introduced in Chapter 2 and may be analyzed in straightforward fashion as described in Chapters 3 and 4 and [237] and [388]. We will refer to them as *standard* bioequivalence cross-over designs. For bioequivalence testing, the same model in SAS is typically utilised to analyse the data as that introduced in Chapter 3 with appropriate modifications for the number of sequences ($s = 1, 2, 3...$), periods ($p = 1, 2, 3...$), and treatments. In SAS, the call to `proc mixed`, the `class` statement, the `model` statement, and the `random` statement are all the same. The `estimate` statement changes to accommodate the comparisons of interest.

An additional alternative design has already been mentioned, the replicate design. It is particularly useful for studying bioequivalence of highly variable drugs. A highly variable drug is defined as a drug with a within-subject CV_W (coefficient of variation) of greater than 30%. The coefficient of variation of a variate is the ratio, expressed as a percentage, of the standard deviation of the variate to its mean. For BE testing we note that $\sigma_W^2 = \ln(CV_W^2 + 1)$, i.e.,

$$CV_W = \sqrt{e^{\sigma_W^2} - 1}.$$

In a replicate cross-over design (see Chapter 4), each subject receives each formulation twice. Eligible subjects are randomized to one of two treatment sequences, e.g., TRTR or RTRT. Thus, each subject is studied in four periods and receives each formulation twice over the course of the study. Similar to the two-period cross-over design described previously, a washout period adequate to the drug under study (at least five half-lives) separates each of the four treatment periods.

Formulation means are estimated with measurement and sampling error in cross-over designs. Replication of measurement within each subject reduces sampling error by a factor equivalent to the number of replications. For example, in a standard cross-over design, the variance of an individual's mean response on $i =$T (or R) is $\sigma_{Bi}^2 + \sigma_{Wi}^2$ where σ_{Bi}^2 is the inter-subject variance and σ_{Wi}^2 is the intra-subject (i.e., sampling error) variance. In a replicate design, the variance of an individual's mean response is $\sigma_{Bi}^2 + (\sigma_{Wi}^2/2)$. Therefore, where high intra-subject variability is of concern, the replicate design will provide more precise estimates of the true individual response.

For a low-variability product, replication does not improve precision dramatically as σ_{Wi}^2 contributes little to the magnitude of the above expression; however, for a high variability product, replication more precisely defines the range over which an individual's mean response may vary. Such measurement is also more accurate as replicate measurement and the derivation of corresponding means converges to the true (and unknown) mean under the central-limit theorem with increasing replication [439]. Such measurement may thus allow for better scrutiny of outliers [458]-[459], but as the comparison of formulation means is of direct concern in the success of average bioequivalence studies, the desirability of such improvement in accuracy and precision is immediately apparent as a practical matter.

The number of subjects required to demonstrate average bioequivalence can be reduced by up to 50% using a replicate design relative to the sample size of a 2×2 cross-over trial. Note that the overall number of doses studied, however, remains similar to a 2×2 cross-over and that the study will be of about twice the duration with twice the blood sampling for each individual subject.

Experience indicates that, although it is theoretically possible to perform a bioequivalence trial with more than four periods, such is rarely utilised. Application of such a trial is not generally limited by logistics (how many subjects can be brought in, length of stay, etc.) but by how much blood can be drawn from a human volunteer in a given time interval! FDA guidance [135] recommends that 12-18 blood samples per subject per period be taken to characterize the PK versus time profile and to derive appropriate estimates of AUC, Cmax, and the other endpoints of interest. A blood collection of 500 mL across the length of a study is the usual limit applied to blood sampling for a human volunteer, and subjects should not have donated blood or plasma for approximately two months prior to being in a study.

Bioequivalence trials must also collect blood samples for purposes other than PK assessment. Such blood sampling for safety assessment (laboratory assessment of liver function, for example) from each volunteer at screening, during the trial, and follow-up in addition to PK sampling limits how much blood can ethically be taken without compromising the safety of volunteers in a given time interval. If this amount is exceeded, this could not only pose a danger to volunteers but also would change the amount of blood available in the circulation and potentially impact the ADME properties of PK measurement (defeating the purpose of collecting it).

We now have an extensive range of options for deciding what type of study design can be applied in a bioequivalence study. These options are applied depending on how many subjects are required to have a good chance of success in demonstrating bioequivalence and the extent of clinical resources. A general algorithm for designing average bioequivalence trials is described below.

Algorithm 5.1: Planning a Bioequivalence Study *[331]*

1. Determine the number of formulations (and doses) to be studied for bioequivalence.
2. Calculate the sample size for a standard cross-over (i.e., a non-replicate 2×2, three or four-period) design. The details of how to perform sample size calculations will be discussed in the next section.
3. Consider available clinical resources.
4. For products with low to moderate intra-subject variation ($CV_W < 30\%$) where adequate resources are available, use the standard cross-over design.
5. For highly variable products, where sample size exceeds available resources, consider a replicate cross-over design and reassess sample size. If

resources are adequate, use the replicate design.

6. For situations where resources are still too limited to achieve desired power, or in situations where one is very uncertain of the magnitude of intra-subject variation (or other critical assumptions), apply a group-sequential design.

Group sequential designs are a further extension of the designs already discussed. These offer the potential for additional resource savings in bioequivalence designs [165], [195]. A group sequential design consists of one or more interim analyses, at which point the sponsor can decide to stop the trial with concrete evidence of success or failure or to carry on. Well known in the statistics community [338], such designs are easy and straightforward to implement in practice in this setting.

A group sequential design approach could be used in cases where there is significant uncertainty about estimates of variability. That is, based on previous data there is a fairly wide range of estimates, such that choosing a lower estimate might result in an underpowered study and choosing a higher estimate might result in an overpowered study, which in either case is a waste of resources. As such, the group sequential design allows one to conduct an interim look with a sample size that provides reasonable power based on a lower (or optimistic) estimate of variability and the final sample size based on a higher (or less optimistic) estimate of variability. Similarly, if uncertainty in the true ratio of bioavailability is of concern, an interim look might be planned based upon sample size required to provide bioequivalence based on the optimistic estimate, with the final look providing conclusive results should this not be the case.

Lastly, a group-sequential design may be applied if it is undesirable to complete a large study due to resource constraints. Some choice of samples for interim analysis may be chosen (based on clinical feasibility) to facilitate an interim look. The probability of success may be quantified at that stage, and if results are inconclusive, the study can continue to completion.

The two aspects of a group sequential design that help determine the probability of stopping early are the alpha-spending function to control the overall Type 1 error rate of the study and the decision rule(s) for stopping at an interim analysis. There are many Type 1 error spending functions and decision rules to chose from, but only those relevant to two-stage group sequential design for a bioequivalence trial will be discussed in this chapter.

The Type 1 error rate as previously discussed was set by regulators at 5% per test for bioequivalence studies and is defined as the probability of a false-positive outcome, or in the case of bioequivalence trails, declaring two formulations are bioequivalent when they are not in truth. Unlike a fixed sample size trial where there is only one analysis, a group sequential trial may have multiple analyses. When data from a fixed sample size trial are analyzed repeatedly during the trial, the overall Type 1 error rate becomes inflated if each look is conducted at the same test level. For example, if two bioequivalence test procedures are conducted (each at the usual 5% level), the overall Type 1 error rate, the probability of a false positive on the first or second test, is 8% (instead of 5%); if three are conducted, the overall rate is 11%; and so on [454].

As such, to control the overall Type 1 error rate of the study, the Type 1 error rate at each analysis must be some value less than the desired overall Type 1 error rate. In a two-stage group sequential bioequivalence trial, the Type 1 error is typically divided equally between the two analyses. A simple, but conservative, method is the Bonferroni adjustment, which results in an error rate of 2.5% per two one-sided test (i.e., 95% CI) at each look, but the resulting overall error rate slightly less than 5%. Another alternative suggested in [343], is to set the error rate at the two analyses at 2.94% (i.e., approximately 94% CI) at each look, resulting in an overall error rate of approximately 5%.

Note that application of such a group-sequential design in bioequivalence testing is not the norm and is in fact discouraged by some guidance [13]. If it is applied, it is expected that most regulators will prefer the position expressed in [13] such that a conservative adjustment (i.e., the Bonferroni procedure) should be applied. For a standard cross-over design with two looks at the data, 95% confidence intervals would be derived at each look.

The decision rule for stopping early (at the first look) should contain both a rule for stopping early when bioequivalence is clearly demonstrated and a rule for futility when bioequivalence is not expected to be demonstrated. For example, one might define the following rule:

1. If the test formulation is demonstrated to be BE at the interim look (i.e., the 95% CIs for AUC and Cmax are contained in (0.80-1.25)), then success has been achieved. Stop the study.

2. If $exp(\hat{\mu}_T - \hat{\mu}_R)$ for AUC or Cmax are not in the range 0.80-1.25, then further study is likely to be futile, and the study should be stopped.

3. Otherwise, continue to the final look.

In the next section, we will consider the calculations which go into deriving a sample size in more detail, and discuss several practical issues impacting the choice of sample size. In the remainder of this section, two other important issues will be discussed: derivation of the variance estimates to be used in bioequivalence sample size calculation, and the length of the washout period.

As discussed in Chapter 1, from the time Phase I starts with the first-time-in-humans study to the time in drug development when one would need to do a bioequivalence study, there is extensive study of pharmacokinetics. In general, AUC, Cmax, and the other PK endpoints are derived in multiple clinical pharmacology studies resulting in a plethora of estimates for σ_W^2 of AUC and Cmax. Each of these study-specific estimates for σ_W^2 may be regarded as independent under the assumption that subjects participating in a given trial do not participate in the other trials.

When independent variability estimates are available across several studies (here studies are denoted $i = 1, 2, ...$), based on the properties of the chi-squared distribution, a method of pooling data across studies is readily available. In brief, a pooled estimate of within-subject variation ($\hat{\sigma}_{PW}^2$) for σ_W^2 is derived as:

$$\hat{\sigma}_{PW}^2 = \frac{\sum_i (n_i - s_i)\hat{\sigma}_{Wi}^2}{\sum_i (n_i - s_i)},$$

where $n_i - s_i$ is the respective df (equal to the sample size n_i less the number of sequences s_i in trial i) for the within-subject variability estimate $\hat{\sigma}_{Wi}^2$ from trial i based on the properties of the chi-squared distribution. These pooled estimates of variation for AUC and Cmax will be utilised in the sample size calculations performed in the next section.

Drugs where no such variability estimates exist for use in calculations are very unusual in the modern pharmaceutical industry, especially for those sponsors who conduct the confirmatory clinical trials and clinical pharmacology programs themselves. Even for drug products where no such in-house data are available, there are a variety of other sources of information (e.g., Physician's Desk Reference, Summary Basis for Approval, etc.) which likely will contain some information on PK variability for use in study design. In the rare event that data are not available, the FDA guidance [135] does make mention of running a pilot study of 6-12 subjects in a cross-over design to estimate variability. Those applying such a technique should ensure that AUC and Cmax data from the pilot study are not pooled with the subsequent confirmatory bioequivalence trial to avoid impacting the Type 1 error rate. Alternatively, a group-sequential analysis plan as described above could be applied.

Estimates of mean and variance for half-life $T_{\frac{1}{2}}$ should also be available across the i trials and can be regarded as independent from one another across trials under the same set of assumptions. One could also pool these estimates to determine an overall mean half-life to define the length of washout period; however, we recommend against such a practice given the importance of an adequate washout in the design of such trials and the interpretation of resulting data.

A key assumption in a cross-over design is that all else being equal, pharmacology and physiology is stable throughout the length of the trial in any given volunteer. This is why normal healthy volunteers are dosed in bioequivalence trials - to prevent the occurrence of or potential for changing disease state from confounding the assessment of any difference in formulations. Therefore, after giving drug and altering (actually causing) PK to be measurable by means of our endpoints AUC and the rest, the washout is utilised to bring back blood concentrations to basal (i.e., null) level, ensure any impact of the drug on the body is negligible, and allow the normal healthy volunteers' bodies to return to 'normal' with respect to blood lost to sampling. They then receive the next formulation in the next period, and so on.

As we have seen in Chapter 3, if concentrations do carry-over through the washout period and into the next period (checked via collection of a blood sample prior to dosing), these carry-over effects confound to some extent the interpretation of differences between formulations. We therefore encourage readers to be over-cautious in the choice of washout period duration in bioequivalence. It should be at least five times the average mean half-life (across studies) and should be extended longer if significant within-subject variation is observed in $T_{\frac{1}{2}}$. Readers interested in a more quantitative definition of the upper limit of mean $T_{\frac{1}{2}}$ may wish to consider the application of a prediction interval (see, for example, Chapter 2 of [314] for more details), but we will not dwell further on this issue here.

For drugs with extremely long half-lives, a parallel group design [135] may be employed. In a parallel group design, subjects are randomized to receive either formulation T (Test) or R (Reference) in a single period and are not crossed over. Such studies are quite unusual in bioequivalence testing and will not be discussed further here. Readers interested in such an approach should see [74] and [447]. Another design in the statistical literature that is sometimes considered is the 'partial'-replicate design. This is simply a replicate design with the fourth period removed (e.g., Examples 3.1.and 3.2). This type of design is seldom applied in average bioequivalence testing; readers interested in using such a design should see [224] and [76] for more information.

5.7 Determining Trial Size

Anyone can run a computer program to calculate how many subjects are needed for a BE study. The actual calculation for determining a sample size is the easiest part of what a statistician does in helping a team design a bioequivalence trial. The calculation itself is straightforward (see Chapter 3, [341], [89], [74], [388], and [237] for background). The more complex part of the job is to ensure that an adequate interface occurs between Statistics and Clinical to ensure the design and sample size are appropriate to the needs of the study.

The first question one should ask Clinical in designing a bioequivalence trial is, ‘How many formulations and doses need to be involved?’ This will help determine the number of periods (2, 3, or 4) to be applied in the study, and thereafter the number of sequences of treatment administration. This number of sequences is critical as it (with n, the number of subjects) will define the degrees of freedom associated with the comparison between formulation means. As sample size is generally small in bioequivalence studies ($n \leq 30$ subjects), an imprecise understanding of the degrees of freedom can lead to an imprecise understanding of the power of a trial and its probability of success.

It is assumed for the purposes of this discussion that within-subject variability estimates are available, for both AUC and Cmax, to determine the trial size. For this purpose the larger of the two pooled estimates is used in calculations, for obvious reasons (i.e., power will be greater, or alternatively the probability of a Type 2 error will be lower, for the endpoint with smaller variation).

The next question to ask Clinical is, ‘How many beds does clinical have, and how many subjects can be scheduled?’ (also known as, ‘How many spots are available?’). Once the extent of those clinical resources has been determined (see Algorithm 5.1), the calculation may be carried out using the SAS code given below to determine power (recall this is 1 minus the probability of a Type 2 error) for the given potential sample size range. Note that this code is general and apples to any standard bioequivalence design, while the code of Chapter 3 is specific in application for only 2 × 2 bioequivalence designs.

```
data a;
* total number of subjects
  (needs to be a multiple of number
  of sequences, seq);
n=30; seq=6;
* significance level;
a=0.05;
* variance of difference of two observations
  on the log scale;
* sigmaW = within-subjects standard deviation;
sigmaW=0.2; s=sqrt(2)*sigmaW;
* error degrees of freedom for cross-over
  with n subjects in total
  assigned equally to seq sequences;
n2=n-seq;
* ratio = mu_T/mu_R;
  ratio=1.00; run;

data b; set a;
* calculate power;
t1=tinv(1-a,n2); t2=-t1;
nc1=(sqrt(n))*((log(ratio)-log(0.8))/s);
nc2=(sqrt(n))*((log(ratio)-log(1.25))/s); df=n2;
prob1=probt(t1,df,nc1); prob2=probt(t2,df,nc2);
answer=prob2-prob1; power=answer*100; run;

proc print data=b; run;
```

For a two-stage group-sequential design (i.e., two looks), the Type 1 error rate (the parameter **a** in the above code) should be reset to 0.025 for the Bonferroni adjustment to determine power for the first look. At the second look, the estimate of the variance (parameter **s**) should also be adjusted for having assessed the variance at the first look in accordance with the findings of [233]. Essentially the variance at the end of the trial is weighted for the relative contribution of degrees of freedom at each look.

We now have a determination of power (probability of success) for our trial for a range of potential sample sizes. However, note that we are ***NOT DONE YET!*** It is important that a sensitivity assessment be carried out to ensure that the power of the trial is not overly influenced by any of our assumptions regarding certain parameters. Additionally, the sample size should ensure that power is sufficient relative to a pre-specified level of random dropouts.

Sensitivity of study outcome to random increases in variability should always be considered by the statistician when providing a sample size estimate. Variation greater than expected will result in less precision (wider than expected confidence intervals) and a drop in power for the study. Some authors [165] recommend derivation of a confidence interval for σ_W^2 for use in sample size calculations with regard to sensitivity analysis, and the authors have found this to be a valuable approach to this issue.

Residual estimates of variability derived from our i studies on the natural logarithmic scale may be considered to be distributed as a chi-squared distribution with degrees of freedom equal to the sum of degrees of freedom associated with the estimates of variability such that

$$\sum_i \frac{(n_i - s_i)\hat{\sigma}_{PW}^2}{\sigma_W^2} \sim \chi^2_{\sum_i n_i - s_i}.$$

Then a (1-α)100% upper confidence bound for σ_W^2 across trials is

$$\hat{\sigma}_{PW}^2 \frac{\sum_i (n_i - s_i)}{\chi^2_{\sum_i n_i - s_i}(\alpha)}$$

where $\chi^2_{\sum_i n_i - s_i}(\alpha)$ is the α quartile of a chi-squared distribution with $\sum_i n_i - s_i$ degrees of freedom.

The next important factor to consider is whether the true bioavailability of the test formulation is the same as the reference formulation. Often this (parameter **ratio** in our code) will randomly differ slightly from unity. Indeed, for highly variable drugs, it is possible for the estimate of the ratio to randomly fall above unity in one trial, and in a follow-up trial of the same formulations to randomly fall below unity! As such, it is not a bad idea to allow for some wobble in the ratio of bioavailability, and FDA guidance generally recommends that sensitivity analyses consider ratios between 0.95 and 1.05.

It is not unusual for subjects to randomly drop out of a trial due to a variety of issues. Food poisoning-induced emesis, the flu, and a family outing are all examples of random reasons why a subject may not participate in a given session of a trial. As the term 'volunteer' implies, subjects have the option to withdraw their consent to participate at any time and are not required to give a reason should they choose not to do so. Such missingness at random in data is easily accommodated by REML modelling but does represent a potential loss in power to the trial as information of such subjects (sometimes termed 'lost to follow up') will not be collected in that period. To compensate, a random dropout rate of 5-10% is generally assumed, and the bioequivalence trial over-enrolls to ensure a sufficient number of subjects complete the trial.

Dosing of subjects at the maximum tolerated dose may also result in dropouts over the course of the study; however it is important to differentiate the 'random' dropouts described above from such a potentially systematic dropout rate related to formulation. If a new formulation is less well tolerated than the reference formulation, this may appear in the data set as an increase in the dropout rate or in adverse event rate on that formulation relative to the reference formulation. Bioequivalence trials are generally not powered to assess the potential for such effects, and we will consider how to assess safety in clinical pharmacology cross-over designs in a later chapter.

One example of this is emesis. Handling of data under this event is treated as a special case in guidance, and may result in data from a subject experiencing this event not being used at all. FDA guidance calls for the following assessment when emesis occurs:

> We recommend that data from subjects who experience emesis during the course of a BE study for immediate-release products be deleted from statistical analysis if vomiting occurs at or before 2 times median Tmax. In the case of modified-release products, the data from subjects who experience emesis any time during the labeled dosing interval can be deleted. [135]

We now turn to an example of determining a sample size. It was required that two new formulations (S and T) be demonstrated as bioequivalent to the clinical trials formulation (denoted R). We planned to use a three-period, six-sequence bioequivalence design and 30-50 spots were expected to be available in the clinical pharmacology unit.

Table 5.6 lists the estimates of within-subject variability available, at that time from previous studies, for use in sample size calculations.

Table 5.6 *Example: Variability Estimates for Use in Designing a Bioequivalence Study*

Study	df	$\hat{\sigma}_W$
1	14	0.23
2	10	0.28
3	10	0.35
4	8	0.15
5	14	0.2
6	24	0.22

Our overall pooled estimate of within-subject standard deviation (σ_W) across studies is $\hat{\sigma}_{PW} = 0.24$ with an upper 50% confidence bound of 0.28. We have at least 30 spots available and as many as 50, and will run the calculations for $n = 30$ and $n = 48$ (recall n must be a multiple of the number of sequences, 6).

Our power for $n = 30$ is 94% and for $n = 48$ is 99%, and under Algorithm 5.1, we conclude that this design will be adequate.

For the sensitivity analysis, we first assess the impact of increased variation up to 0.28 standard deviations. Power for $n = 30$ is 82% and for $n = 48$ is 97%. The authors' rule of thumb is that at least 80% power should be maintained under random changes in assumptions.

Second we assess the impact of a change in relative bioavailability to 0.95 instead of unity. Power for $n = 30$ is 85% and for $n = 48$ is 96%.

We assess the impact on power if we are very unlucky and variation increases to 0.28 standard deviations along with a decrease in true bioavailability to 0.95. Power for $n = 30$ is 73% and for $n = 48$ is 90%. This last scenario is pretty unlikely, but it does not hurt to check.

Last, the drug was well tolerated at the maximum dose, so over-enrollment on the order of 5% is likely called for, so we would over-enroll one or two subjects to ensure that the minimum desired number complete the study.

In this situation, n of 30-36 subjects should provide at least 90% power (most likely) and probably at least 80% power if the assumptions are not too grossly violated.

5.8 What Outliers Are and How to Handle Their Data

Although not explicitly stated in regulatory bioequivalence guidance (see Chapter 2), there is a very great distinction between an outlier in a statistical sense and in a regulatory sense.

In statistical training, outliers are generally introduced as a topic related to assessment of model fit (see for example, [314] Chapters 2 and 5). An outlier is defined as a residual that has large value - i.e., the model does not fit the data point well. Various statistical procedures and tests have been devised over the years to identify such outliers. In general, if the absolute residual value corrected for variation (termed 'studentised residual') is greater than 2 or 3 (depending on how conservative one is), then a data point may be termed a *statistical* outlier.

In terms of statistical impact, an outlier (or set of outliers) may impact the estimate $\hat{\delta}$ (by influencing its position relative to 0) and inflate the estimate of within-subject variance $\hat{\sigma}^2_W$ (resulting in a wider than expected confidence interval) or both. Impacting either of these parameters implicitly makes it more difficult to demonstrate average bioequivalence. The previous section provides a quantitative means of determining the potential impact on power (the probability of a successful BE trial).

This statistical assessment is a purely quantitative approach - providing an objective assessment of whether or not a data point is unusual. Statistical science rarely goes further - i.e., to describe what should be done with such data points. In reporting of BE trials, it is usual for statistical results to be presented with and without the outlying data points to determine if the outlier is influential on the results. This, however, leaves one in a practical quandary - which analysis is to be believed?

Outliers may not be excluded from a bioequivalence data set on statistical grounds alone. From a regulatory review perspective, handling of such data is very difficult, and the FDA and other regulatory agencies require that such data be looked at quite carefully.

If an outlier is the result of a protocol deviation (for example the subject drank far too much water or chewed up the pill instead of just swallowing it), then deleting the outlier from the data set may be justified [135]. However, if evidence of such a deviation does not exist, regulators assume that the cause of the outlier is either product failure (maybe the tablet dissolved in some strange manner) or due to a subject-by-formulation interaction (for example, the new formulation might be more bioavailable than the reference formulation in certain subjects). Admittedly, it may also just be a random event, and there is generally no way to differentiate between which of the three categories a given outlier belongs. Average bioequivalence studies are not designed to assess individual product differences but only to compare the formulations means.

In this context, whether an outlier is a product failure, a subject-by-formulation interaction, or a random event is immaterial. These are confounded, and final inference with regard to bioequivalence (and regulatory approval) is based on the full data set (i.e., including the outliers). If observations are deleted, it is the sponsor's responsibility to provide a rationale to convince the regulators that such is appropriate.

On a practical level, this essentially means in practice that there is no such thing as an outlier in a bioequivalence data set, and while we recommend that statisticians always check the assumptions of their model, in this context there is little utility in spending too much time worrying about outliers' impact on the findings. The authors have never seen an instance where a protocol violation has resulted in regulators deeming the deletion of a data point as acceptable; however, the authors have seen many instances where outliers from one to two subjects have resulted in a conclusion that bioequivalence was not demonstrated. Example 3.2 (Chapter 3) could be viewed as an example of this. From the data it was impossible to rule out that product failure, a subject-by-formulation interaction, or just a random occurrence of an outlying data point were involved. This generally results in a follow-up BE trial (of similar design) being done as was described in the prologue to this chapter.

As the impact of outliers cannot be controlled after the study completes, the best way to deal with them is to acknowledge that they can happen at random and to protect the study's power for random appearance of outliers at the design stage. To do so, it is recommended that bioequivalence studies be powered at 90% and that such trials have at least 80% power under potential inflation of the variability estimate and for potential changes in δ of up to 5%.

5.9 Bayesian BE Assessment

The approach to statistics thus far described is deductive in that we collect data to test a specific hypothesis or set of hypotheses. We use *observed* data to derive statistics to test specific facts in which we are interested. Statistics are used to quantify the probability that the data collected are consistent with our predetermined hypotheses.

For bioequivalence testing, one could denote the probability of observed data given the hypotheses conditions as:

$$p(\underline{y}|H_{01}, H_{02}). \tag{5.1}$$

where $\underline{y}$ denotes the observed data, and H_{01}, H_{02} refer to the two one-sided hypotheses of interest for the difference between formulation means (see Equations (2.3) and (2.4)). This direct, deductive approach to statistics enjoys a very long history [176], and is the most often referred to approach in biopharmaceutical statistics. It is referred to as direct probability assessment as it deals 'directly' with the observed data to draw conclusions.

However, this is not the only way to consider looking at data. One might approach data in an inductive manner. In this case, we have a predetermined (rough) idea of the state of nature, and we collect data to give us a more precise idea of this state. This approach is inductive in that we *assume* we know what is happening or will happen, and data are collected to reinforce the point. This approach to statistics also enjoys a long history and was developed in the late 1700s and early 1800s by Bayes and Laplace [176]. It is referred to as indirect probability assessment as it deals 'indirectly' with the observed data to draw conclusions about the unknown parameters of interest. In average bioequivalence testing, for example, $\delta = \mu_T - \mu_R$, the true (unknown) difference in formulation means is the parameter of most interest.

In biopharmaceutical statistics, indirect probability assessment is not employed as often as direct probability assessment. The reason is quite obvious, given a brief rereading of Chapter 1. When making regulatory claims, it is the sponsor's burden to prove to a regulator's satisfaction that the drug is safe, efficacious, and can be manufactured to good quality. The regulator's presumption is that it is not safe, efficacious, nor of good quality until sufficient data are provided to prove otherwise. Therefore, the application of indirect probability assessment is of limited practical utility in regulated bioequivalence testing. Indirect probability assessment is also deemed subjective in that one must make assumptions regarding the conditions being studied.

Note, however, that when one is not working in a confirmatory setting but is *exploring* clinical development of a compound (as described in Chapter 1), an inductive approach to statistics adds much value. Under such circumstances, the sponsor is working under the assumption that a drug is reasonably safe, efficacious, and of good quality and wishes to collect data to design and study the drug in confirmatory clinical trials. For example, one might perform trials in clinical development to determine the best tolerated and effective dose to subsequently be used in a confirmatory trial. In such a setting, it is not necessary to demonstrate to regulators that such is the case, but only to provide sufficient evidence to satisfy internal decision makers in the sponsoring company. Use of Bayesian inference offers substantial benefit in terms of data exploration (see [42]). Such an inductive approach will be discussed further in Chapters 7 and 8. Here we will include a brief discussion on indirect probability assessment in bioequivalence for completeness and to introduce concepts to be used later.

In mathematical terms, we first acknowledge we have some idea of what is happening or will happen with the difference in formulation means (expressed as $p(\delta)$). We again collect data from the BE study and calculate a probability; however, here we are interested in the probability of the conditions for δ given the observed data, which we denote as:

$$p(\delta|\underline{y})$$

rather than the probability of observing the data given a hypothetical set of conditions. Note there is no explicitly stated hypothesis in this indirect probability setting.

The derivation of an indirect probability may be extremely complex mathematically, and this was a practical bar to the implementation of such approaches to data analysis until recently. Modern computing software has rendered this complexity manageable. Recent developments in Markov-Chain-Monte-Carlo-based methods known as Gibbs sampling (e.g.,WINBUGS at http://www.mrc-bsu.cam.ac.uk/bugs/) were developed in the late 1980s and 1990s [161] to implement indirect probability assessment in a straightforward fashion. We will now consider an example in bioequivalence to illustrate the concepts; further illustration of these methods for normal data models may be found in [156].

One first assumes a functional form for $p(\delta)$ and then derives

$$p(\delta|\underline{y}) = \frac{p(\underline{y}|\delta)p(\delta)}{p(\underline{y})}. \tag{5.2}$$

This expression indicates that the probability in which we are interested $(p(\delta|\underline{y}))$ is equal to the probability of the data given the possible values of $\mu_T - \mu_R = \delta$ multiplied by our initial idea about the properties of the situation $(p(\delta))$ divided by the overall probability that one would observe the data $(p(\underline{y}))$. Note that the expression $p(\underline{y}|\delta)$ is very similar to Equation (5.1). Under certain conditions, inductive and deductive reasoning will yield similar findings statistically, and we will observe such in this example.

We now turn to the example utilizing the AUC and Cmax data found in Example 3.1. WINBUGS code for this analysis is provided in the Technical Appendix.

This model assumes we know very little about the true values of the unknown parameters of interest (here, $\mu_T - \mu_R$). Here, for example, the WINBUGS parameter `delta` (δ in our mathematical notation) is assumed to be normally distributed with mean 0 and an very wide standard deviation of 1000. This is termed a noninformative (i.e., flat) prior distribution. WINBUGS then uses a Gibbs sampler (100,000 iterations were performed for this example) according to the procedure developed in [156] to integrate the data and derive the probability distribution for $p(\delta|\underline{y})$ and $p(\sigma^2_W|\underline{y})$. From the estimated distribution, we can derive various statistics. In Table 5.7 we choose to present medians and 90% confidence intervals. The median is a statistic derived by taking the value in the distribution where 50% of data falls above and 50% of data falls below the value. In keeping with the findings of Laplace [176], the posterior median is the most appropriate measure of centrality for a distribution when using indirect probability assessment.

The findings of Table 5.7 should seem very familiar. If we review these findings relative to the analyses of Chapter 3, one will find that they are very similar. This is generally the case if one uses a noninformative prior distribution as the weight introduced by this term into Equation (5.2) is minor.

The key problem with the regulatory application of indirect probability assessment in this setting is this dependence on the assumptions of the prior distributions. For example, if one makes a minor change to the model (assumes, for example, that `delta` has a distribution with different moments, such as a standard deviation of 1) the distribution of $p(\delta|\underline{y})$ may sometimes change dramatically. The extent of such a change depends upon the weight of this term in Equation (5.2) relative to the

Table 5.7 *Statistics for δ and σ_W^2 Inverse Probabilities Given AUC and Cmax Data Observed in Example 3.1*

Parameter	Median δ (90% BCI)	median σ_W^2
AUC	-0.0166 (-0.0614, 0.0286)	0.0114
Cmax	-0.0270 (-0.1121, 0.0588)	0.0410

BCI = Bayesian Confidence Interval
5th and 95th Quartiles of p given observed data

weight of the observed data. In this example, such is not the case, but changes in the assumptions for smaller data sets can result in a different inference, complicating regulatory interpretation when using such a method [329]-[330]. We leave sensitivity analysis to assess this potential to the reader. Code for the examination of data in Example 3.2 may be found on the website accompanying this book.

5.10 Technical Appendix

5.10.1 SAS Code for Simulating BE Designs

The following SAS code may be used to create 5000 simulated logAUC data for 5000 subjects, for a randomized cross-over trial where subjects are switched from the reference formulation to the test formulation and vice-versa.

The first `data` statement sets the parameters to be used in the simulations, and then derives variance-covariance estimates for the data (see Chapter 3 for details). The code then selects random numbers for a random variable `t` based on the seed value `123`. Unless a seed is specified, SAS will use the clock time to derive the random number. A similar procedure is followed for the random variable `r` accounting for the fact that `r` will be correlated to `t` in line with the findings of Chapter 3. These random numbers are combined with the parameter values to create the 'Monte Carlo' observations.

SAS Simulation Code:

```
data simtr simrt;
  keep subject t_pk r_pk;
    mut=180;*mean of test;
    mur=200;*mean of reference;
    varb=0.18;*between-subject variance;
    varw=0.09;*within-subject variance;
    per1=0;*period effect in period 1;
    per2=0;*period effect in period 2;
    seqtr=0;*sequence effect in TR;
    seqrt=0;*sequence effect in RT;

    var=varb+varw; rho=varb/var;
    std=sqrt(var);
    c=sqrt(1-rho**2);

    do i = 1 to 4999 by 2;
       t = rannor(123);
       r = rho*t+c*rannor(123);
       t_pk = seqtr + per1 + mut + std*t;
       r_pk = seqtr + per2 + mur + std*r;
        subject=i;
         output simtr;end;

    do i = 2 to 5000 by 2;
       t = rannor(456);
       r = rho*t+c*rannor(456);
       t_pk = seqrt + per2 + mut + std*t;
       r_pk = seqrt + per1 + mur + std*r;
        subject=i;
         output simrt;end;
run;

data simtr;set simtr;sequence='TR';run;

data simrt;set simrt;sequence='RT';run;

data simulate;set simtr simrt; run;
```

5.10.2 Bayesian Statistical Theory in BE

Consideration of methods for BE decision making in the early 1980s focused on the use of Bayesian posterior probabilities for the construction of comparisons for μ_T:μ_R. Rodda and Davis [364] and Mandallaz and Mau [289] evaluated the decision rules introduced in [451]-[452] and introduced consideration of this distribution relative to a predetermined goalpost interval of (-Δ, Δ); bioequivalence was concluded if the posterior probability of falling in this interval was higher than a predetermined probability level, e.g. 0.9.

This idea was further developed in [148] which introduced graphical methods to accompany the consideration of the posterior probability and recommended that Δ be altered according to the drug under study. Selwyn et al. [381] and Grieve [170] developed methods for the Bayesian analysis of the randomized, two-period cross-over and evaluated the impact of various other factors (carry-over, choice of prior distributions) on inference. Reisner and Guttman [357] developed similar ideas in the engineering field, and Yee [466] developed a non-Bayesian method for deriving the upper and lower bounds of the probabilities for rejecting bioequivalence.

In summary, these methods used the normal and gamma distributions and the models of Chapter 3 to derive probabilistic statements using Bayes' rule on the posterior probability for the difference of μ_T-μ_R given the data observed to assess the degree of average bioequivalence. Between and within-subject variances were assumed to be independent, and fixed effects were assumed to be normally distributed with mean and nested variance appropriate to the model.

Prior distributions have to be specified for all model parameters. Nuisance effects (period effects) are integrated out of the log-likelihood function using an appropriate method based on Bayes' function (full details may be found in [381] and [170]). The posterior distribution for $\mu_T - \mu_R$ is based on the prior distributions, model, and data which is calculated using Bayes' rule [273].

Numerical integration or approximate methods [381]-[382] were initially proposed for use in implementing these techniques; however, these were known to be subject to various problems (e.g., impact of starting values and sensitivity to numerical assumptions, computer intensive), and while the techniques offered substantial benefit in the practical assessment of bioequivalence, their use was not encouraged by regulators within industry applications [329]-[330].

Use of a Bayesian procedure was known to be potentially sensitive to the choice of prior distribution - a classic topic of debate between Bayesians (those in favor of indirect probability assessment) and Frequentists (those in favor of direct probability assessment). This led to questionable validity in implementation in public health. Whether this situation will hold true in the future (as such methods become more and more used in drug development) is open to debate.

It should be noted that a Bayesian analysis naturally facilitates the use of sequential experimentation to assess bioequivalence. Extensions to the Bayesian approach by first conducting a pilot relative bioavailability study to estimate within-subject variability in a two-step procedure were discussed in [351]. A small pilot study (sample size of six subjects) is first conducted under this approach for the purpose of deriving prior beliefs (or distributions). A full-size bioequivalence study is then conducted based on this information to assess bioequivalence under predetermined Regulatory standards. Other Bayesian approaches to the assessment of ratios of means are described in [16]. More recent developments may be found in [159].

Example 3.1 AUC and Cmax data were separately analyzed in WINBUGS using the following computer code (based on the code from [237] Chapter 2):

```
model;
{
for( k in 1 : P )
{
for( j in 1 : N )
{
Y[j , k] ∼ dnorm(m[j , k], precw)
m[j , k] < − mu + subject[j]
+ equals(k,1)*((pi/2) + seq[j]*delta/2)
+ equals(k,2)*(-(pi/2) - seq[j]*delta/2)
} }
for( j in 1 : N )
{
subject[j] ∼ dnorm(0.0, precb)
}
precb ∼ dgamma(0.001, 0.001)
precw ∼ dgamma(0.001, 0.001)
mu ∼ dnorm(0.0, 1.0E-6)
pi ∼ dnorm(0.0, 1.0E-6)
delta ∼ dnorm(0.0, 1.0E-6)
sigma2w < − 1 / precw
sigma2b < − 1 / precb
theta < − exp(delta)
}
```

Subjects' (j = 1 to N) log-transformed observations (in periods k=1 or 2) are deemed to be normally distributed with mean (m[j , k]) and inverse within subject variance (precw). The mean m[j , k] is a function of the overall mean (mu, included to center the model on the overall average between and within subjects), each subject as their own control (subject[j] assumed to have inverse variance precb), and period (pi) and formulation (delta). In the first period, m[j , 1] = mu + subject[j] + (pi/2) + (seq[j]*delta/2) where seq[j] indicates to WINBUGS whether the subject (j) received the Test or Reference formulation in the first period based on the sequence of treatments to which each subject j was randomized. In the second period, m[j , 2] = mu + subject[j] -(pi/2) - (seq[j]*delta/2).

Note the change in sign (+ to −) associated with the terms for period and formulation effects. This indicates that the period and formulation effects in period 2 should be the opposite of period 1, and as the model is centered on the overall mean we must divide pi and delta by two to indicate to WINBUGS that for any given subject the distance between an individual subject's two observations due to period and formulation effects should equal `pi` and `delta`, respectively.

CHAPTER 6

The Future and Recent Past of BE Testing

Introduction

A few years later, I was asked to attend a meeting in Hilton Head, South Carolina, where bioequivalence was one of the topics of discussion. There were presentations by several statisticians from the FDA, academia, and industry on the topic. I regarded this as somewhat of a pain - there was a lot of work to do, I had a date that weekend, and I could not see where flying off to Hilton Head was going to be helpful to anyone at all.

My boss, however, vetoed my not going. It was expected that I would attend (and eventually participate in) such conferences as a matter of professional development, representing the company and the discipline of statistics (etc., etc.). Also, she did not have time to go. So I dutifully packed my bags and headed down. One of the reasons I had gone to work was that I was tired of sitting through lectures, but I left secure in the knowledge that at least maybe I could possibly play golf while down there.

When the conference was over, I came back and reported on the upcoming new FDA proposals about assessing bioequivalence (to be discussed later in this chapter). I was still pretty new to the company and industry at this point, so how bioequivalence testing was done did not really bother me one way or the other. As long as I knew what to do with the data and how to design the studies, I was holding up my end. The FDA was planning to issue a draft guidance on the topic later that year.

The reaction I received was kind of like the reaction one gets when accidently knocking over a bees' nest - the bees are very surprised, kind of annoyed, do not like it, and may be less than friendly. My boss was very surprised by the information I brought back, and to be blunt, did not believe me. I argued about it with her for a while, showed her my notes, and pointed out that if she did not like the message, it was her fault as she was the one who had made me go. That was not helpful in resolving the argument. In the end, I had her invite one of the local academic statisticians who had given a talk at the meeting to come to 'the Unit' to discuss the upcoming FDA proposals.

If she did not believe me, I figured she would believe him. Either way, it *was fine by me. I had work to do, and it was windy and rained the whole time, so I had not gotten to play golf. My going to the conference had been a disappointment to everyone as far as I could tell, and I resolved to do so as little as possible in future (little did I know....).*

It is amazing how often this type of thing happens in industry (not the arguments - that happens every day - the inviting of external people to make a point). I have had to do this type of thing several times since then. You may know exactly what is going on for a particular issue, but very often people at the company want to hear it themselves from someone else external to the company before they will believe that they really have to do anything about it. It has been pointed out that we have to pay these people to come talk to us (i.e., this approach is not really cost-effective), but that is how business is often done.

After the external academician came in and spoke with us, my boss believed me, and there was a great deal of discussion at the company about the possible implications of this proposal (nobody knew) and when it would come into effect (no one knew that either). In the end, my boss asked that I go down to Washington, DC, with her the following winter after the draft guidance was issued [122] for a special FDA Advisory Board meeting on the topic. These are meetings of experts (external to the FDA) on a particular topic who advise the FDA on how to protect public health.

In the end, this resulted in my spending the next approximately five years working on this area of bioequivalence, doing extensive research and presenting at various meetings here, there, and everywhere on the topic and its implications for public health. It was important and also interesting research, and I saw most of the airports in North America (and beyond).

The lesson of this experience is:

1. *Conference attendance is actually important. It keeps one on the cutting edge at work.*
2. *In the modern world, it is not enough to just do your day job. Working folks should engage in research that benefits them professionally at their company and also externally.*
3. *All that said, five years of research is a long time and a lot of research. Be careful what conferences you choose to attend, and never go to Washington, DC, for an FDA meeting with your boss if you can help it.*

6.1 Brief History

The FDA proposed the use of individual (IBE) and population (PBE) bioequivalence as a method for ensuring bioequivalence of new formulations in 1997 [122]. These methods (defined later in this chapter) were to replace the use of average bioequivalence (ABE) already discussed in previous chapters.

This generated a great deal of discussion and public debate, and the IBE and PBE proposals were amended in 1999, [125]-[126], and finalized in 2000-2001, [130]-[131]. After FDA reviewed data from application of such techniques in practice, the IBE and PBE methods were removed from their guidance in 2003 [135].

We discuss the IBE and PBE history and application here for completeness; however, these approaches to bioequivalence should **not** be used in regulatory submissions. We will discuss the difficulties in implementation of these approaches later in the chapter.

First, some background. Sheiner [401], Schall and Luus [371], and Schall [370] introduced an alternative method for bioequivalence assessment based on models of dose-response [399], risk assessment, and different combinations of estimates of means and variances from models. Under such a moment-based approach to bioequivalence assessment, differences in means and variances are combined into one aggregate statistic for the assessment of population and individual bioequivalence. If the upper 95% bound on the aggregate statistic falls below a pre-set equivalence margin, bioequivalence was demonstrated. Such a procedure implicitly would allow for widening (or narrowing) of the equivalence margin based upon variation observed in the study.

Bootstrap [370] or Bayesian [401] (see Chapter 5) based assessment of the quartiles of the aggregate endpoint were initially proposed; however, estimation procedures for such a statistic using approximation procedures involving the Cornish-Fisher Expansion [31] and methods for the linear combination of independently Chi-squared distributed variables ([216]; [144]; [214]; [180]; [53]; [169]; [284]; [428]; [440]; [52]) were developed in more detail in [209]-[210].

Practical strategies for population and individual bioequivalence assessment under this approach were developed in [372], and the application to the moment-based criterion of most interest to the FDA was developed in greater detail by Hyslop et al., [223]-[224]. Alternative parametric procedures were described in [194], [249], [76] and [350].

It should be noted that many other statistical approaches were considered during the debate on bioequivalence. Testing procedures for assessing differences in means and variances simultaneously (though not as a composite endpoint) were developed in [21], [22], [65], and [158]. Stepwise procedures (testing for equivalence in means between formulations followed by testing for equivalence in variances) were described in [103],[104], [437], [436], [173], [174], and [167]. Unbiased, optimal tests for bioequivalence assessment were described in [312], [215], [48], [441],

[442] and multivariate, optimal assessment of bioequivalence (e.g, for AUC and Cmax simultaneously) were described in [28], [29], [72], and [313]. Testing for differences in profiles was described in [294].

The US Food and Drug Administration's decision following the debate on whether population and individual bioequivalence were needed to protect public health and the approach chosen for assessment were announced in draft guidance released in 1997 [122] based on the principles discussed in [372]. Previously discussed approaches to moment-based assessment of population and individual bioequivalence were established as described in later paragraphs for studies conducted prior to approval and following approval of new chemical entities.

Average bioequivalence was deemed insufficient to protect the public health as it assessed only the difference in formulation means, did not adjust for the variance of narrow therapeutic drug products and highly variable drug products, and did not account for assessment of subject-by-formulation interaction. However, no evidence of therapeutic failure had been established over the five years in which the 1992 FDA guidance had been in effect [17].

Conventional two-period, randomized, well-controlled, cross-over designs were established as the design to be performed in the assessment of population bioequivalence for approval of bioequivalence in formulation changes prior to approval of the new drug product [122]. However, two-sequence, four-period (RTRT, TRTR), randomized, well-controlled, *replicate* cross-over designs (described in Chapter 4) were chosen as the design to be performed in the assessment of individual bioequivalence for approval of new formulations following approval of the new drug product for both generic manufacturers and those manufacturers wishing to make formulation changes following approval. Replicate designs were required for the assessment of individual bioequivalence so that within-subject estimates of variance were estimable along with the subject-by-formulation interaction [122].

Overall, the FDA preliminary draft guidance [122] involved little change in study design for sponsors conducting trials to establish bioequivalence of a new commercial formulation relative to that used in clinical trials under the population bioequivalence approach to inference, though different analyses were recommended for data analysis and decision making.

The new draft FDA guidance, however, required replicate designs for changes in formulations following approval - a more complex design for the majority of drug products. Under this approach to inference, logAUC and logCmax were to be analyzed separately using a two-stage (REML) linear model including terms for sequence, period, and formulation in the model for a replicate design. Within-subject variability estimates were to be derived for each formulation, and subject-by-formulation interaction was to be used to assess whether there were subgroups of subjects responding differently to the new formulation.

We now turn to the statistics used for this purpose by the FDA. Following this discussion, we turn to a topic currently being debated - scaled average bioequivalence.

6.2 Individual and Population BE

Population bioequivalence was to be assessed using the following aggregate statistic [122]:

$$\frac{(\mu_T - \mu_R)^2 + \sigma_T^2 - \sigma_R^2}{max(0.04, \sigma_R^2)}, \tag{6.1}$$

where $\sigma_T^2 = \sigma_{WT}^2 + \sigma_{BT}^2$ and $\sigma_R^2 = \sigma_{WR}^2 + \sigma_{BR}^2$. Note that this aggregate statistic can be constructed using a mixed model from a two-period cross-over design or a parallel group design and does not require the use of a replicate design.

Individual bioequivalence was to be assessed using the following aggregate statistic [122]:

$$\frac{(\mu_T - \mu_R)^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2}{max(0.04, \sigma_{WR}^2)}. \tag{6.2}$$

Because the within-subject variance of each formulation cannot be separately estimated from between-subject variance estimates in most two-period cross-over designs of the form (TR, RT), a replicate design is generally required.

The goalpost for population bioequivalence assessment assumed a total-subject variance for the reference formulation of 0.04 and was set to approximately 1.74 as follows:

$$\frac{(\ln(1.25))^2 + (0.02)}{0.04}, \tag{6.3}$$

allowing for a mean difference of 20% on the log_e-scale and a variance allowance of 0.02 in the numerator under the procedure proposed by the FDA [122]. If the upper 95% percent bound on the FDA metric fell below this value, population bioequivalence was demonstrated for the endpoint under study. The goalpost for individual bioequivalence assessment assumed a within-subject variance for the reference formulation of 0.04 and was set to approximately 2.49 as follows:

$$\frac{(\ln(1.25))^2 + (0.03) + (0.02)}{0.04}, \tag{6.4}$$

allowing for a mean difference of 20% and a variance allowance of 0.03 in the numerator for subject-by-formulation interaction and 0.02 for the difference in within-subject variance under the procedure proposed by the FDA [122]. If the upper 95% bound on the FDA metric fell below this value of 2.49, individual bioequivalence was demonstrated for the endpoint under study.

To undertake these assessments at least 1500 (2000 bootstrap samples were recommended in [122] and [96]), while preserving the number of subjects in each sequence and a mixed model appropriate to the design was to be fitted to each bootstrap sample (see Chapter 5). The appropriate aggregate statistic, either (6.1) or (6.2), was then to be derived based on the model estimates for each bootstrap sample; note that the denominator for each bootstrap's aggregate statistic was to be chosen based on the point estimate from the model estimates of the original data set. The nonparametric percentile method [96] was then to be used to calculate an upper 95% bound for the quantity of interest. It was required that the upper 95% bound for the metric of interest fall below predetermined regulatory bounds (1.74 and 2.49 for population and individual bioequivalence, respectively) for both AUC and Cmax for bioequivalence to have been demonstrated.

Bootstrap-based inference was deemed undesirable as it introduces randomness (see Chapter 5) and because the coverage probability of the nonparametric percentile method was observed to fall below regulatory standards [335]. Hyslop et al., [223]-[224], developed a parametric approach to inference using method-of-moments estimates which was later supplemented by an asymptotic testing procedure for situations where missing data were present [237]. We will confine discussion of the statistics involved to IBE for the purposes of this chapter. Information on PBE may be found in [332], [335], and [237].

Hyslop et al.'s [223]-[224] method first calls for the linearization of the statistic of interest. This simplifies the approach to hypothesis testing and is simply a matter of mathematical convenience. Under this approach, the hypotheses of interest for IBE become:

$$H_{0R} : \delta^2 + \sigma_D^2 + \sigma_{WT}^2 - (3.49)\sigma_{WR}^2 \geq 0 \tag{6.5}$$

if $\hat{\sigma}_{WR} \geq 0.2$. This is referred to as the reference (product variation) metric, and

$$H_{0C} : \delta^2 + \sigma_D^2 + \sigma_{WT}^2 - \sigma_{WR}^2 - (0.996) \geq 0 \tag{6.6}$$

if $\hat{\sigma}_{WR} < 0.2$ (a constant scaled metric).

Note that the hypotheses H_{0R} and H_{0C} are not continuous at the break-point $\hat{\sigma}_{WR} = 0.2$, and this led to issues with conflicting findings in certain settings [473]. FDA guidance suggests using both constant and reference scaled metrics under certain circumstances:

> VII. D. Discontinuity: The mixed-scaling approach has a discontinuity at the changeover point, sW0 (individual BE criterion) or sT0 (population BE

> criterion), from constant- to reference-scaling. For example, if the estimate of the within-subject standard deviation of the reference is just above the changeover point, the confidence interval will be wider than just below. In this context, the confidence interval could pass the predetermined BE limit if the estimate is just below the boundary and could fail if just above. This guidance recommends that sponsors applying the individual BE approach may use either reference-scaling or constant-scaling at either side of the changeover point. With this approach, the multiple testing inflates the type I error rate slightly, to approximately 6.5%, but only over a small interval of sWR (about 0.18-0.20).[131]

In practice, therefore, those testing for IBE were to derive both tests, and if $\hat{\sigma}_{WR}$ was near 0.2 could use rejection of either test as sufficient evidence to demonstrate IBE.

Method-of-moments estimates for the moments of interest may be placed in this expression and the Cornish-Fisher expansion [235] may be applied to calculate an approximate upper bound [223]-[224] for complete data sets using confidence bounds for each of the parameters as follows.

For the linearized version of the FDA's IBE metric, a procedure is described in the FDA guidance [131] based on [223]-[224] that is appropriate for replicate cross-over designs with no missing data and is summarized as follows. In this situation, the estimates $\hat{\delta}$, $\hat{\sigma}^2_I$, $\hat{\sigma}^2_{WT}$, and $\hat{\sigma}^2_{WR}$ are derived for $\delta = \mu_T - \mu_R$, $\sigma^2_I = \sigma^2_D + \frac{\sigma^2_{WT}+\sigma^2_{WR}}{2}$, σ^2_{WT}, and σ^2_{WR} based on method-of-moment estimates and utilised as follows:

1. Derive unbiased, independent method-of-moments estimators $\hat{\delta}$, $\hat{\sigma}^2_I$, $\hat{\sigma}^2_{WT}$ and $\hat{\sigma}^2_{WR}$.
2. Let H_δ be the square of the absolute value of the larger of the lower and upper 90% bounds on δ derived using the t-distribution and using the Satterthwaite approximation [369] for the degrees of freedom, $H_I = \frac{\nu(\hat{\sigma}^2_I)}{\chi^2_\nu(0.05)}$, $H_T = \frac{\nu(\hat{\sigma}^2_{WT})}{2\chi^2_\nu(0.05)}$ $H_R = \frac{-(3.99)\nu(\hat{\sigma}^2_{WR})}{\chi^2_\nu(0.95)}$ where $\chi^2_\nu(\alpha)$ is the αth-percentile point of the chi-squared distribution with ν degrees of freedom.
3. Then
$$(\hat{\delta}^2 + \hat{\sigma}^2_I + \frac{\hat{\sigma}^2_{WT}}{2} - (3.99)\hat{\sigma}^2_{WR})$$
$$\mp[(H_\delta - \hat{\delta}^2)^2 + (H_I - \hat{\sigma}^2_I)^2 + (H_T - \frac{\hat{\sigma}^2_T}{2})^2 + (H_R - (-3.99\hat{\sigma}^2_{WR}))^2]^{\frac{1}{2}}$$
is an approximate, 90% confidence interval for the FDA's IBE metric.

Appropriate modifications to this approach were to be made when the metric is scaled to a constant variance and for population bioequivalence assessment (see [131]).

Alternatively, one may have used an asymptotic test, [332], [335], as described in [237] or use a modified bootstrap procedure, [395]-[396], to perform these tests.

Table 6.1 *IBE Derived Upper 95% Bounds for Chapter 5 Replicate Design Examples*

Endpoint	$\hat{\sigma}_{WR}$	REF Hys	CST Hys	REF Asy	REF Asy	REF Bot	CST Bot
`exam5.sd2`							
AUC	0.080	0.008	-0.079	0.006	-0.084	0.008	-0.085
Cmax	0.209	0.012	-0.003	0.009	-0.035	0.003	-0.034
`exam6.sd2`							
AUC	0.344	-0.195	-0.065	-0.179	-0.067	-0.009	0.020
Cmax	0.557	-0.223	0.306	-0.169	0.290	-0.065	0.345

REF: Reference-Scaled Hypothesis
CST: Constant-Scaled Hypothesis
Hys=Hyslop; Asy=Asymptotic; Bot=Bootstrap

These calculations are quite complex. Given the limited utility of application of these procedures in regulatory application, we conclude discussion of the statistics here and will only note (for interested readers) the following results for examples of replicate designs with data in Chapter 5.

Both data sets would have been deemed IBE as the upper bound of interest falls below zero. Note the discontinuity findings for `exam5.sd2`, Cmax. As one of the tests falls below zero and the $\hat{\sigma}_{WR}$ lies very near 0.20, IBE was demonstrated. Also note that the test formulation is IBE to the test formulation for Cmax of `exam6.sd2`. As we know from Chapter 5, this test failed to demonstrate ABE due to a dramatic, statistically significant increase in mean exposure for the test relative to the reference formulation. In this case, within-subject variation is so large ($\hat{\sigma}_{WR}$=0.557) as to overwhelm the difference in means in the test for IBE. In this instance, a formulation that was not ABE could be declared IBE and presumably allowed access to market.

To call something equivalent implies a context or criteria for the determination. There are several stakeholders in determining such a criteria:

1. Statistical considerations: the approach should be quantifiable, accurate, precise, well understood, and should be transparent in interpretation.

2. Sponsor considerations: Using a well-designed, controlled, and rea-

sonably sized study (or set of studies), the sponsor should be able to show the criteria have been met with a quantified chance of success.

3. Regulatory and public-health considerations: The approach used must protect public health (in that the risk of false positive market access must be controlled at a predetermined rate).

Statistically, the IBE approach could be accommodated. The Hyslop et al., [223]-[224], asymptotic [237], and bootstrap tests, [122], [395]-[396], were available to perform the tests of interest.

The interests of sponsors were also addressed in the FDA's [131] final statistical guidance. The procedures could be carried out using small, well-controlled cross-over designs.

However, the regulatory application of these techniques was found not to protect public health for several reasons:

1. The property of the IBE procedure to allow more than a 20% change in average relative bioavailability for highly variable drugs was deemed unacceptable. Hauck et al. [197] demonstrated that a widening from the traditional average bioequivalence 0.80-1.25 acceptance criteria could lead to therapeutic failure in the marketplace.
2. The weighting of the IBE findings to observed reference product variation was thought to be unacceptable. The more variable this finding, the easier it was to demonstrate bioequivalence under the IBE and PBE approaches. Inflation of this variance by running a poorly controlled design could improve the chances of successfully demonstrating bioequivalence.
3. The interpretation of variance estimates being plugged-into the IBE statistic was questionable. The variance estimates could become biased by their estimation method, [109]-[110], and in the presence of missing data [335]. Variance estimates from such small studies are poorly characterized statistically; however, these variances play a key role in the success or failure of an individual bioequivalence trial, having far greater weight than the difference in formulations means, [472]-[473].
4. While reasons 1 to 3 above might have been sufficient to prevent any new formulation from entering the market using such approaches, the issue was exacerbated when multiple new formulations entered the marketplace at patent expiration under the IBE approach to testing. IBE did not protect public health in that dramatic changes in average exposure were possible when multiple formulations entered the market at patent expiration [335]. As patients would be switching from formulation in an uncontrolled manner, dependent on which formulation they picked up at the pharmacy, the potential for widespread therapeutic failure was too great to allow for this potential application. The

traditional ABE testing prevents this possibility and protects public health [10].

In the end, FDA guidance [135] withdrew the possibility of application of IBE and PBE due to these and other reasons. The problem remains as to how to demonstrate average bioequivalence for highly variable drugs, though as we have seen previously, replicate and group-sequential designs may be used for this purpose. Another alternative, sometimes debated for highly variable drugs, is a statistical procedure known as Scaled Average Bioequivalence (SABE), and we briefly develop the properties of such a procedure in the next section.

6.3 Scaled Average BE

Highly variable drugs, and also narrow therapeutic index drugs (see Chapter 2), have been the subject of debate at various times as it was felt that the 0.80–1.25 acceptance range might not protect public health (in the latter case) and was too stringent (in the former case), requiring large trials.

Tothfalusi and Endrenyi [430] provide an excellent review of the topic, expanding on their consideration of the topic in [429]. In essence, scaled average bioequivalence may be viewed as a special case of individual bioequivalence where $\sigma_{WT}^2 = \sigma_{WR}^2 = \sigma_W^2$ and $\sigma_D^2 = 0$.

As such, the statistic of interest becomes:

$$\frac{(\mu_T - \mu_R)^2}{\sigma_W^2}. \tag{6.7}$$

Note that in the denominator of this expression, $max(0.04)$ is not included, as in IBE and PBE testing. This is in keeping with the description of [430]. In recognition of the FDA's determination that narrow therapeutic index drugs need not be held to a strict standard of bioequivalence [135], we include here the fixed scaling parameter 0.04 to protect narrow therapeutic index drugs from an overly stringent bioequivalence standard.

For drugs with low variability ($\sigma_W^2 \leq 0.04$), the traditional average bioequivalence tests (see previous chapters) are used [430]. For larger variation, the two one-sided tests become:

$$H_{01}: \quad \frac{\mu_T - \mu_R}{\sigma_W} \leq -\eta \tag{6.8}$$

versus the alternative

$$H_{11}: \quad \frac{\mu_T - \mu_R}{\sigma_W} > -\eta$$

and

$$H_{02}: \quad \frac{\mu_T - \mu_R}{\sigma_W} \geq \eta \tag{6.9}$$

versus the alternative

$$H_{12}: \quad \frac{\mu_T - \mu_R}{\sigma_W} < \eta.$$

As with ABE, in a 2 × 2 cross-over, this two one-sided test procedure may be assessed using a confidence interval. Tothfalusi and Endrenyi [430] stated that:

$$[t_{0.05}(\lambda, n-2), t_{0.95}(\lambda, n-2)]$$

is a 90% confidence interval for $\frac{\mu_T - \mu_R}{\sigma_W}$ where t_α denotes the α quartile of a noncentral t distribution with noncentrality parameter $\lambda = \frac{(\hat{\mu}_T - \hat{\mu}_R)\sqrt{n/2}}{\hat{\sigma}_W}$ with $n-2$ degrees of freedom.

If these limits lie between $-\eta\sqrt{n/2}$ and $\eta\sqrt{n/2}$, then scaled average bioequivalence is demonstrated.

Consider Example 3.1 from Chapter 3. The statistics of interest may be derived by entering the appropriate values into the following SAS code. We utilize $\eta = 0.795$ for the purposes of this example as discussed in [430].

```
data sabeauc;
eta=0.795;n=32;d=-0.01655; s2=0.01100;
lambda=d/(s2**0.5);
t_05=TINV(0.05,30,lambda); t_95=TINV(0.95,30,lambda);
ll=-eta*((n/2)**(0.5)); ul=eta*((n/2)**(0.5)); run;

proc print data=sabeauc noobs;
var ll t_05 t_95 ul;run;

data sabecmax;
eta=0.795;n=32;d=-0.02694; s2=0.03835;
lambda=d/(s2**0.5);
t_05=TINV(0.05,30,lambda);t_95=TINV(0.95,30,lambda);
ll=-eta*((n/2)**(0.5));ul=eta*((n/2)**(0.5)); run;

proc print data=sabecamx noobs;
var ll t_05 t_95 ul;run;
```

In this analysis of Example 3.1, the lower limits of interest are -1.86 and -1.84, and the upper limits are 1.53 and 1.55 for AUC and Cmax respectively, indicating that scaled average bioequivalence was demonstrated as these fall within the limits (-3.18 and 3.18).

Code is not provided for Example 3.2; however, interested readers may use the findings of Chapter 3 and code similar to the above to determine that, in Example 3.2, scaled average bioequivalence was also demonstrated. Note that in Example 3.2, average bioequivalence (the traditional approach) was not demonstrated.

As an exact procedure is available, extension of the Hyslop et al. [223]-[224] procedure and the application of asymptotic and bootstrap testing are not considered here. Interested readers should see [429] for more information on extension of the Hyslop et al., [223]-[224] procedure as applied to scaled average bioequivalence testing.

Values of η discussed in the literature range from ± 0.7 to ± 1.1 versus the traditional average bioequivalence limits of $\pm \ln 1.25 = \pm 0.223$. Regulatory agencies have yet to define an appropriate value of η in relevant guidance, and it is unlikely that they will do so.

The reasons for this follow from the debate around IBE and PBE. Expansion of the acceptance limits is unacceptable from a regulatory perspective as such a procedure would not protect public health [197]. As with IBE and PBE, acceptance criteria as described above would to some extent still be design dependent - in that running a ‘poor’ study (i.e., a poorly controlled study) would yield high variability making it easier to demonstrate scaled average bioequivalence.

Regulatory acceptance of scaled average bioequivalence is unknown at present but unlikely in the future, and this scaled average bioequivalence procedure should not be used in regulatory applications at the present time. Given the extensive debate around bioequivalence testing in the recent past, average bioequivalence may be expected to be the standard procedure for some time to come.

CHAPTER 7

Clinical Pharmacology Safety Studies

Introduction

One day, out of seemingly nowhere, I received a very strange request from a clinical scientist. We will call her Betty, and she asked if I could round off a confidence interval? My immediate response was, 'No. Why would anyone want to do that?'

In essence, we had derived an upper bound in a drug interaction trial of 1.2538 for AUC. Evaluation of this value relative to the acceptance level of 1.25 showed that it was higher than 1.25. We could not conclude the two treatments were equivalent. Pretty elementary. Betty wanted to round it off, so she could claim equivalence had been demonstrated.

I told her no, and left it at that. Such would misrepresent the data, and the statistics underlying the upper bound could not support 'rounding it off'. Clearly as the value was higher than 1.25, the null hypothesis had not been rejected, and it was out of the realm of possibility. To my mind it was also a matter of professional integrity, and I was a bit surprised that anyone would ask such a thing. The less I said, the better off we would both be.

However, I was still new on the job, and did not know that some people will not take no for an answer, even if it is a matter of professional integrity. So began one of my most important 'learning experiences' on the job. 'Learning experiences' are a business euphemism for an experience no one in their right mind wants any part of, but you are stuck with it because you work there.

Rounding off turned out to be really, very important to Betty and the physician for whom she worked, and a major disagreement at the company developed. Peoples' egos became involved, and everyone who had even only a nebulous stake in this (or a potentially related) issue felt compelled to comment. Academic experts were paid and consulted. Opinions were sought from the FDA on the topic. Many internal meetings on the topic were held, and (despite their best efforts to avoid it) several senior vice presidents had to be consulted and in the end backed us up, 'No rounding'.

Years later FDA guidance [135] was issued saying the same thing, but *as is often the case, such business precedes regulatory guidance by many years.*

Guess who was at the center of this argument? It was a rough experience (for what I still feel was a ridiculous request), but I learned a lot from interacting with such people on such a thing and from watching how they and many other people behaved. If had it to do over again, I would have followed a different approach to dealing with such people. I call it the 'Nurse' approach in honor of the people whom I saw do it.

We had a drug intended for the treatment of hypertension (high blood pressure) which caused migraines if given at high doses. We discovered this in the first study in man (which is designed for this purpose, see Section 7.1), and carefully worked out at which dose the problem started. These were bad migraines - the throwing-up kind. The study team wanted to stop the study, but a chief medic said to continue. The rationale was that they wanted to explore more doses before going to the next study.

There was no point in continuing. The study had defined the maximum tolerated dose, completing its objective. We were at an impasse with the medic involved. We discussed the ethical issue of continuing (i.e., not), but were told headache and emesis were not a serious enough side effect to warrant not exploring further. Egos began to become involved. Senior vice presidents were again getting phone calls.

This came to an abrupt stop, and the nurses put a stop to it. I am told that they told chief medic that if he wanted to continue, he'd have to come down and clean up the vomit himself. The study ended the next day. That was not the official reason logged in the study file, and it is hearsay, but I think it is probably true.

The moral of the story is that when you are asked to do something you consider inappropriate, put the person who is asking in your shoes. When they will actually have to get their own hands (or shoes) dirty to do such a thing and take personal accountability for it, you will be surprised at how the pressure to do so suddenly lets up. If not, then try 'No'.

When exploring safety, it is important that we get it right for the sake of each and every patient who will take the product. Everyone has a stake in this assessment. Even the people who develop and sell drugs may themselves have to take them one day! All drugs have side-effects and should be presumed to be unsafe if used incorrectly. Some side-effects can be very serious and life-threatening.

The role of clinical pharmacology safety studies is to define how the body handles the drug such that side-effects can be predicted in a rational, scientific manner. This assessment determines how the drug should be used correctly to treat the condition under study. Every decimal point matters. Do not cut any corners which would compromise patients' safety, *and ensure your findings represent the data accurately, so that the people using the drug can make a fully informed decision.*

7.1 Background

All other things considered, it is comparatively easy to tell when a drug is efficacious. The drug should change something about the body or its characteristics for the better, making people live longer or healthier or both. A drug that does not offer such benefit (referred to as medical utility or efficacy, see Chapter 2) would presumably not be approved for sale to a human population. The problem in drug development is to detect, observe, and ensure that the change is to the benefit of patients.

Drugs that are unsafe, producing unwanted, nonbeneficial side-effects presumably should not be approved. This, however, constitutes a more complex issue (and one that is still evolving). The difficulty is how to deduce how and when a drug is safe. In contrast to efficacy assessment, in safety assessment the problem is to assess and ensure that no clinically relevant change in the health status of patients results from use of the drug beyond the decreased health status associated with natural factors (like aging, for example).

This problem initially seems similar to bioequivalence testing in that the desired outcome is to test for no change in the potential for hazard relative to control agent (say another drug in the same class) or placebo. The problem is different in that in bioequivalence testing, we understand and have a historical basis for the assessment of the potential for hazard using pharmacokinetics as a biomarker - i.e., if AUC goes down too much in a new formulation, efficacy may be lost, and if Cmax goes up too much, side-effects may appear.

In safety testing for new drugs, though, we do not know what the potential for hazard actually is in a human population! The relationship of rate and extent of exposure needs to be established relative to unknown (but presumably present) side-effects before such an assessment is valid scientifically.

Our working assumption is initially that the drug is not safe when given at any dose in any formulation under any circumstances to any human population. As a practical matter, it is also important to recognize that we will never be able to demonstrate the alternative to this assumption - i.e., that the drug is safe at any dose in any formulation under all circumstances when given to any person. All drugs are potentially toxic if used incorrectly; however, some may be used at appropriate, carefully selected, and studied doses in controlled circumstances to treat diseases in particular populations.

Following preclinical safety assessment to ensure that the new drug is not toxic at low doses (discussed in greater detail in the next section), clinical pharmacology safety assessment of a new drug product usually starts with giving the drug in very low doses and placebo to a robust, healthy population - normal healthy volunteers. The rationale for doing so is that is that if the new drug causes unexpected side-effects, healthy people are most likely to recover. It is relatively easy to monitor them closely, and any side effects identified will not be confounded with disease (as normal healthy volunteers should not have any). Some patients may eventually be willing to tolerate side-effects if their underlying disease is treatable, but one cannot really assess that potential until one knows what the side-effects are! Sometimes, however, it is impossible to dose normal healthy volunteers (e.g., it is unethical to give a cytotoxic oncology agent to a normal healthy person). For such drugs, clinical pharmacology safety assessment begins in patients with the condition under study.

Dosing starts with very low doses, well under the no adverse effect level (NOAEL) seen in the most sensitive preclinical species, and slowly the dose is increased in these initial safety studies until:

1. Side-effects are observed (e.g., nausea, headache, changes in laboratory values), or
2. Rate (Cmax) or extent (AUC) of exposure approach the NOAEL.

The intent of these small (generally cross-over [293]), well-controlled, cautious designs is to carefully assess evidence of the potential of the drug to cause a hazard to people taking the drug. Note, however, that absence of evidence is NOT evidence of absence [238]. If side-effects are not observed and dosing is halted with exposures near the NOAEL, the potential for significant hazard still exists (even if remote). If a side-effect is observed, its relationship to exposure and dose may then be quantified. Additionally, once a potential hazard is identified, safety may be assessed relative to other agents used for treating the population for which the drug is intended.

The role of statistics in this setting is different from bioequivalence testing. Here, statistics are used to quantify the unknown relationship of unwanted side-effects to dose and exposure while dose is varied over the course of the study. A non-null relationship of dose or exposure to a safety endpoint demonstrates the *statistical* potential for hazard [225]. Note however that statistical potential does not necessarily imply that the drug is unsafe and should not be used or developed. Its benefits (efficacy) may outweigh the presence of these side-effects, but that is up to the clinicians, regulators, and patients who will be using the drug to determine. Statistics provide an impartial assessment in this setting to aid them in making this determination. All drugs are unsafe; some are useful under carefully controlled circumstances (to limit the risks involved).

Once the relationship of dose and exposure to safety is understood, clinical pharmacology studies are then performed to assess under what circumstances it is safe to administer the drug. For example, one would study what happens when the drug is given with and without food or with and without another drug.

In this chapter, we will explore commonly used statistical methods for clinical pharmacology assessment of dose and assessments of certain circumstances to determine if and when the drug can be dosed with a reasonable expectation of safety while treating a disease. Such studies limit but do NOT eliminate the potential for hazard when using a drug. Such cannot be eliminated with 100% certainty as we know from Chapter 1.

Note that 'reasonable expectation' is not well defined in regulatory guidance. Safety is currently an emerging scientific topic (e.g., [229]). Whether a drug is safe (or not) is subjective. Physicians, patients, regulators, and drug-makers all have different opinions on the topic.

Operationally, and usually, statistics are derived posthoc - after the study has ended. Decisions about what dose to give in these studies and how to dose titrate are made by clinical personnel. The role of statistics is to assess and precisely quantify the relationship of dose to pharmacokinetics and dose to safety endpoints once the study completes.

In some situations [399], [402], [403], [320], [328], quantitative interactive models may be used to assist clinical personnel in selecting doses 'on-line', and we will consider an example in next section. This is by no means an exhaustive list of work on this topic, and readers may wish to examine recent publications on the topic (e.g., [427], [94], and [456]). These procedures use models to predict what effects will be observed at different doses to aid in clinical decision making. However, the final decision about what dose is used is the physician's responsibility and the subject's or patient's responsibility before taking a drug.

Safety assessment in clinical pharmacology and drug development is a rapidly evolving science. It has not always done well in the rush to get drugs to needy patient populations. Historically, [426] over-dosing is common as a result, and there have been numerous circumstances where the approved dose of drug has been reduced once the drug has been on the market for a time. It has been said [339] that in the 20th century drugs were presumed to be safe until shown otherwise (this is hard to believe in a litigious society). However, increasing attention from regulatory agencies is being applied to this area in light of recent safety risks, and major refinements and improvements in how we test for safety in clinical studies may be expected in the coming years.

7.2 First-time-in-humans

The administration of a drug to humans for the first time generates a great deal of excitement in the sponsoring organization and is an exciting time for everyone involved. New therapies offer potential benefit to numerous patients. Before such a drug can be administered, however, it must undergo an extensive battery of in vitro and in vivo preclinical testing. In certain nations (e.g., USA, Europe), first-time-in-humans study protocols and their supporting preclinical information must also be submitted to and approved by the relevant regulatory authorities (see [120] for an example).

Regulators will, in general, desire to review the following items prior to administration of a new drug to humans [120]:

1. The First-time-in-humans (FTiH) study protocol,
2. Information on the chemistry, manufacturing, and control/stability of the drug manufacturing process,
3. information on preclinical Pharmacology and Toxicology in vitro and in vivo studies (containing, at a minimum, an integrated summary of animal toxicology study findings and the study protocols), and
4. Any human experience with the investigational drug (e.g., if studies were carried out in a different nation).

We now turn to consideration of the FTiH trial design, conduct, and analysis and will not discuss these regulatory requirements further here.

In contrast to bioequivalence trials (the objective of which is to confirm equivalence of different formulations), the objective of FTiH and other Phase I trials is to learn [402] about the safety, pharmacokinetic, and pharmacodynamic properties of the drug being studied. The application of statistics to this topic of drug development is fundamentally different from that used in the confirmatory setting of bioequivalence, though the study designs, conduct, and models used in such studies are similar.

The approach to data analysis and interpretation is typically inductive (see Chapter 5) in that those performing FTiH and Phase 1 studies have a 'rough' idea of how the drug will behave (from the preclinical testing described previously). Studies are performed and data are collected to reinforce this 'rough' idea. The role of statistics in this setting is to employ the tools discussed previously (Chapter 1: randomisation, replication, blinding, blocking, and modelling) to ensure the estimates provided by such studies are accurate and precise.

It would be desirable if the preclinical findings were perfectly predictive of what one would observe in humans for a new drug, but this is not always the case. There are interspecies differences which preclude such a possibility (for example, see Chapter 31 [12]). George Box stated that, 'To find out what happens to a system when you interfere with it you have to interfere with it (not just observe it).' [39], and the assumption made for any new drug is that it will cause undesirable side-effects (hereafter referred to as adverse events, AEs) that are dose and exposure dependent, in that the higher the dose or exposure, the more likely such an AE will occur.

An adverse experience (AE) is any untoward medical occurrence in a patient or clinical investigation subject, temporally associated with the use of a medicinal product, whether or not considered related to the medicinal product. Such events are frequently characterized as:

1. Mild: An event that is easily tolerated by the subject, causing minimal discomfort and not interfering with everyday activities.
2. Moderate: An event that is sufficiently discomforting to interfere with normal everyday activities.
3. Severe: An event that prevents normal everyday activities.

A Severe AE is an AE that is noticed and alarming (e.g., severe nausea or emesis), but does not necessarily require cessation of treatment (the disease under study, like cancer, might make such an event tolerable though undesirable).

In contrast, a **Serious** AE (SAE) is 'an event that is fatal, life-threatening, requires in-patient hospitalization or prolongs hospitalization, results in persistent or significant disability, or results in congenital anomaly or birth defect' (Chapter 14 [37]). Observation of such an SAE in a FTiH study would generally halt dosing for all subjects being studied and must be reported quickly to relevant regulatory authorities.

In FTiH trials, dose is increased as knowledge is gained of the drug's properties until a 'potential for hazard' is observed. 'Potential for hazard' in this context denotes the observation of conditions where it is possible for an adverse reaction to drug treatment to occur or the actual observation of a serious or severe AE. A dose just lower than this dose is defined as the maximum tolerated dose (MTD) [225].

Statistical proof of hazard (i.e., a p-value less than 0.05 for a comparison of $H_0 : \mu_D - \mu_P \leq 0$ where μ_D denotes the mean effect at a dose and μ_P denotes the mean effect on placebo [199]) may or may not be obtained in such studies. Determination of the MTD is often driven more by clinical judgment and less by statistical analysis given the limited numbers of subjects exposed to drug in such studies. If the drug-induced rate of an adverse experience in the population is p for particular dose, then the chance one sees at least one such event in n subjects exposed to a dose of drug in a study is $1 - (1 - p)^n$. As FTiH trials typically involve only a small number of patients or subjects (sample sizes per dose ranging from $n = 6$ to 10), p must be relatively large in order to observe an AE in the trial. For example, if $p = 0.1$ (the proportion of subjects experiencing for example a headache caused by drug at a dose) and $n = 6$ subjects are studied at this dose, the probability of observing at least one subject with a headache in the trial is only 0.47 at this dose.

For a rare side-effect (drug-induced neutropenia, for example), with a $p = 0.01$, the probability of observing such an event in a FTiH trial is only 0.06 with $n = 6$. Thus FTiH trials are geared toward detection of non-rare side-effects. If the drug causes a side-effect in less than 5% to 10% of people at a given dose, it is most unlikely that such trials will observe such an event.

Cross-over designs are generally employed for the purposes of informative dose-escalation in FTiH and Phase I studies as such designs are known to be more informative and provide better information than alternative designs [399] and expose only a limited number of subjects to the (potentially) harmful agent. See Table 7.1, for example. Dosing is conducted in separate cohorts, sequentially, with results from each dose being reviewed prior to the next dose being administered in the next period. Periods are separated by a washout sufficient to ensure no drug is on board when the next dose is given (generally at least one week to allow for pharmacokinetic washout and review of data).

Placebo is administered to serve as a control for evaluation of any AEs observed, and subjects are randomly assigned to the period in which they receive it. Subjects are generally kept blinded as to whether they have received drug or placebo in order to ensure this assessment is unbiased by knowledge of treatment.

Depending on the NOAEL and properties of the drug under study, shorter cross-over designs (i.e., two-period or three-period designs) may be employed. For particularly toxic drugs, oncology trials of cytotoxic agents are generally conducted using a parallel group design where cohorts of patients are randomized to increasing doses of drug (Chapter 1 [37]). We will consider an example later in this chapter but will first focus attention on how to model data from a typical trial. Such techniques also apply to the shorter cross-over designs described above.

Preclinical pharmacology and toxicology data are used to choose the FTiH starting doses. The preclinical pharmacology and toxicology studies should identify a no-effect dose and a no-adverse-effect exposure level in multiple pre-clinical species. Allometric scaling ([356]; [133]; Chapter 8 [37]) is then applied to estimate a safe starting dose. In essence, allometric scaling uses the NOAEL and accounts for differences in weight and physiology between species to yield a range of doses expected to

Table 7.1 *Schematic Plan of a First-time-in-humans Cross-over Study*

Subject	Period 1	Period 2	Period 3	Period 4
Cohort 1				
1	P	D1	D2	D3
2	D1	P	D2	D3
3	D1	D2	P	D3
4	D1	D2	D3	P
Cohort 2				
5	P	D4	D5	D6
6	D4	P	D5	D6
7	D4	D5	P	D6
8	D4	D5	D6	P
Cohort 3				
9	P	D7	D8	D9
......				

P=Placebo; D1=Lowest Dose
D2=2nd lowest dose; etc.

be safe in humans. The NOAEL in the most sensitive species (i.e., the lowest NOAEL) is defined as the upper limit of human exposure (AUC and Cmax, as previously).

Once a presumed safe range of doses is estimated, an algebraic dose escalation scheme (1x, 2x, 3x, 4x, etc.), geometric dose escalation scheme (1x, 2x, 4x, 8x, etc.), or Fibonacci scheme (Chapter 1 [37] and Chapter 31 [12]) are used to determine the next dose to administer in the next period or cohort of subjects. The choice of dose escalation scheme is prespecified in the study protocol. The choice of next dose may be reduced (but not increased) relative to the intended, protocol-specified, scheme dependent on the results from the previous dose.

Subjects or patients participating in FTiH studies are monitored very closely for the occurrence of AEs. Subjects are generally required to stay in bed for at least 4 hours following a dose, and continuous monitoring of vital signs is not unusual for a period of at least 24 hours following each dose. The population enrolled into a FTiH study is generally composed of male healthy volunteers as females are known to be more prone to drug-induced toxicity [306]. Full discussion on inclusion and exclusion criteria for subjects enrolled in FTiH trials may be found in Chapter 1 [37] and Chapter 31 [12] and will not be discussed further here.

Operationally, each cohort of subjects is brought into a clinic on a weekly basis. Following an overnight fast, the dose chosen (or placebo) is administered at roughly 8 a.m., and safety, pharmacokinetic, and pharmacodynamic (if any) measurements are taken prior to dosing and at regular intervals thereafter. These data are then used by the study team (composed at minimum of a physician, nurse, statistician, and pharmacokineticist) to support the decision on which dose to give next (or whether to halt or delay the next administration). The key responsibility for determination of which dose to administer next (if any) is a medical purview, and the statistician and pharmacokineticist are expected to provide analyses and simulations to support this medical determination if required. The statistical and pharmacostatistical approach to data analysis in this setting is exploratory (see Chapter 14 [37]). Data are modelled, periodically during the study, to provide an accurate and precise description of what observations have been collected to date and are used to predict which effects may be observed at future doses ([177]; [1]; chapter 18 [12]).

We first consider pharmacokinetic data generated in a typical FTiH trial. One property of such log-normal pharmacokinetic data is that variation increases with exposure [453]. To model this behavior, a 'power' model is generally utilised [408]. Doses are increased until average exposure (AUC and/or Cmax) is observed to approach the NOAEL or some multiple of the NOAEL's value (e.g., one-tenth). For this type of design, the power model is:

$$y_{ik} = (\alpha + \xi_k) + \beta(ld) + \varepsilon_{ik},$$

where α is the overall mean pharmacokinetic response at a unit dose (logDose, $ld = 0$) known in statistics as the population intercept, ξ_k is the random-intercept accounting for each subject (k) as their own control, β is the slope parameter of interest regressed on logDose (parameter ld), and ε_{ik} denotes within-subject error as described in Chapter 3 for each log-transformed AUC or Cmax (y_{ik}) in period i. Note that period effects are assumed to be minor relative to the magnitude of effect of logDose in this analysis and are confounded with dose. Typical data arising from such a design are listed in Table 7.2 and plotted in Figure 7.1.

Table 7.2: Example 7.2.1: AUC and Cmax Data from a Cross-over First-time-in-humans Study Design

Subject	Period	Dose	AUC	Cmax
1	2	15	666.06	307.1

Table 7.2: Example 7.2.1: AUC and Cmax Data from a Cross-over First-time-in-humans Study Design

Subject	Period	Dose	AUC	Cmax
1	3	45	1701.49	524.2
1	4	100	4291.86	1684.2
2	1	5	144.63	70.1
2	3	45	956.84	390.9
2	4	100	2121.55	522.0
3	1	5	187.88	55.6
3	2	15	406.06	210.1
3	4	100	2712.69	864.6
4	1	5	111.12	53.7
4	2	15	313.21	155.8
4	3	45	1006.57	548.7
6	1	5	152.64	96.3
6	3	45	1164.88	520.7
6	4	100	3025.78	1509.1
7	2	15	641.89	233.6
7	3	45	2582.20	713.0
7	4	100	4836.58	1583.7
8	1	5	420.42	212.7
8	2	15	908.93	339.3
8	4	100	8194.40	2767.2
9	1	100	3544.28	947.0
9	2	150	5298.14	778.9
9	3	200	6936.13	1424.4
10	1	100	5051.23	1713.3
10	3	200	11881.12	3543.8
10	4	250	16409.81	4610.1
12	2	150	7460.82	2143.2
12	3	200	8995.97	3708.4
12	4	250	10479.14	2604.0
14	1	100	2134.17	1664.5
14	2	150	3294.38	932.4
14	4	250	5332.19	1276.3
15	2	150	3189.74	976.2
15	3	200	4643.52	1300.7
15	4	250	4652.96	810.1
16	1	100	3357.67	1134.8
16	2	150	4305.17	856.8
16	3	200	8886.62	1914.2
17	1	5	378.75	155.1

Table 7.2: Example 7.2.1: AUC and Cmax Data from a Cross-over First-time-in-humans Study Design

Subject	Period	Dose	AUC	Cmax
17	2	15	915.95	307.2
17	3	45	2830.42	532.8
18	1	100	1912.93	596.3
18	2	150	2684.00	602.6
18	4	250	3971.27	1792.2
19	1	100	8446.20	2110.6
19	3	200	17004.51	2766.3
19	4	250	21097.81	7313.4

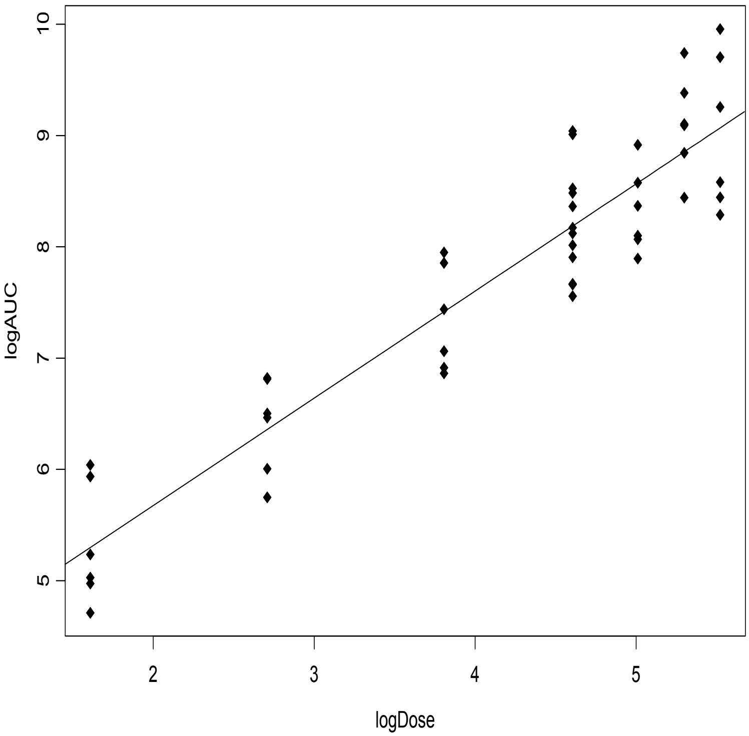

Figure 7.1 *Estimated logDose versus logAUC Curve with Individual Data Points from Example 7.2.1*

Note that variation at the 150 mg dose in Example 7.2.1 (see Figure

7.1) appears to decrease relative to the 100 mg dose. This is a feature of the cross-over nature of the design and is observed due to the fact that the subjects administered the 150 mg dose are not always the same ones administered the 100 mg dose. To account for each subject as their own control, the power model is utilised to provide a population dose to pharmacokinetic response curve. This statistical relationship provides an estimate of the magnitude of a typical individual's exposure when administered a dose. Once a subject's exposure has been measured for a given dose, this individual's dose to pharmacokinetic relationship may be quantified to provide an individual assessment of potential hazard relative to the NOAEL, and we will consider how to do so later in this chapter. SAS code to model AUC and Cmax data from such trials is below. Doses are increased until the population dose to pharmacokinetic curve approaches the NOAEL or until a severe or serious AE is observed.

First-time-in-humans PK SAS `proc mixed` *Analysis Code Example 7.2.1 and 7.2.2*

```
proc mixed method=reml data=pk1_ftih;
    class subject;
    model lnauc=lndose/
    s ddfm=kenwardroger cl alpha=.1;
    random intercept/subject=subject;
    run;
```

SAS `proc mixed` output provides the estimates required to derive the dose to AUC or Cmax curve plotted in Figure 7.1. SAS output (not shown) estimates of parameters are given in Table 7.3.

Table 7.3 *Parameter Estimates from Example 7.2.1*

Endpoint	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2_W$
AUC	3.75	0.96	0.01
Cmax	3.20	0.83	0.09

The parameter α in this example is the estimated logAUC (or logCmax) associated with a dose of 1 mg ($ld = 0$). The estimated population dose to pharmacokinetic response curve is calculated as:

$$AUC = e^{\hat{\alpha}+\hat{\beta}(ld)}.$$

To solve for the dose expected to yield exposure at the NOAEL (the

MTD), one exponentiates the above equation at $AUC = NOAEL$ after solving for ld:

$$MTD = e^{\frac{\ln(NOAEL)-\hat{\alpha}}{\hat{\beta}}}.$$

The bootstrap (see Chapter 5 and [179]) may be used to derive a confidence interval for the MTD if desired.

In our second example (Example 7.2.2, see Table 7.4) we consider a PK data set where exposure relative to a predetermined NOAEL was of concern. Dosing was to be halted if mean AUC was in excess of 2400 ng.h/mL or Cmax exceeded 880 ng/mL (the NOAEL).

Table 7.4: Example 7.2.2: AUC and Cmax Data from a Cross-over First-time-in-humans Study Design

Subject	Dose	AUC	Cmax
1	1	611	80.3
1	5	842	103.1
1	10	1600	167.3
2	1	1052	112.7
2	5	1584	164.7
2	10	2809	273.8
3	1	1139	98.0
3	5	1896	162.6
3	10	2531	167.9
4	1	989	89.0
4	5	1604	177.6
4	10	1817	212.8
5	1	1275	114.2
5	5	2282	173.7
6	1	947	77.7
6	5	1698	138.0
6	10	2278	240.5
7	1	603	92.3
7	5	1289	149.5
7	10	1987	225.5
8	1	867	86.4
8	5	1263	130.7
8	10	2494	276.3

Estimates of the parameters of interest may be found in Table 7.5. Here it was observed that exposure approached the NOAEL for AUC at the 10 mg dose and dosing was halted accordingly. See Figure 7.2.

Individual fitted means at each dose with 90% confidence intervals may be derived easily in SAS `proc mixed`. A statement `outp=pred` is

Table 7.5 *Parameter Estimates from Example 7.2.2*

Endpoint	$\hat{\beta}$	$\hat{\sigma}_W^2$
AUC	0.38	0.02
Cmax	0.36	0.03

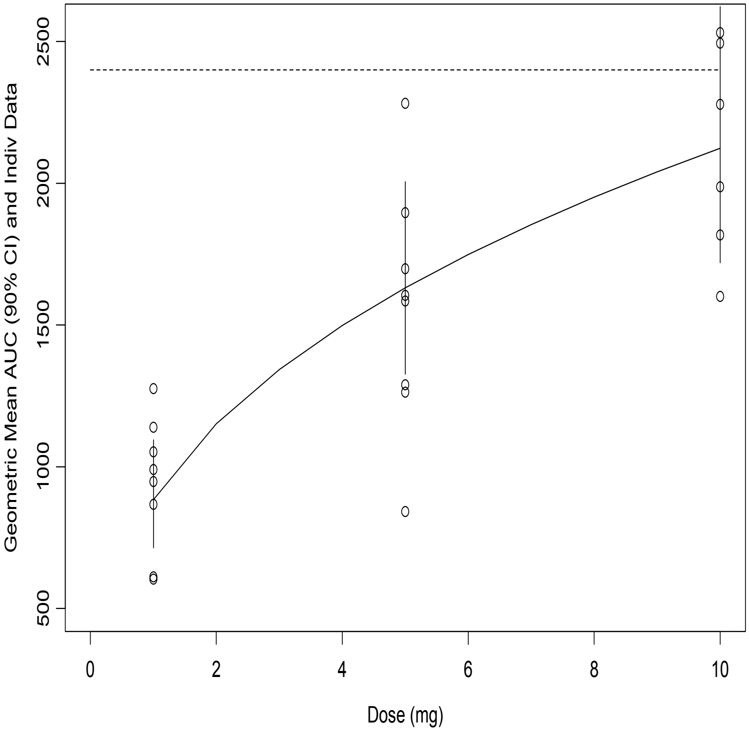

Figure 7.2 *Estimated Dose versus AUC Curve (90% CI) with Individual Data Points from Example 7.2.2*

added to the model statement after the / to output the data set `pred` containing the relevant values. Estimated responses at other doses may be obtained by entering a missing value for the observation desired for that subject. Code to perform such analyses are provided on the website accompanying this book, and consideration is left as an exercise for interested readers.

In normal healthy volunteer studies, severe AEs are unusual, and SAEs are very unusual. Observation of an SAE should halt all dosing in a study and requires regulatory scrutiny of the event. Dose escalation is halted if severe AEs are observed. However, it is unusual for either SAEs or severe AEs to be observed in such trials. Most often dose escalation is halted when mean exposure approaches the NOAEL (as seen in the example above). Dosing for any given individual is halted if their exposure data approaches a higher than expected factor of the NOAEL.

In contrast, FTiH studies for cytotoxic agents are performed in refractory patient populations, and the goal of the study is to identify a dose causing a dose-limiting toxicity (DLT, an SAE) with X% frequency (often 30%). This is referred to as the dose expected to cause an X% response, abbreviated ED_X. The assumption is that for such an agent to be efficacious, it must approach toxic levels. Three patients are dosed with a low dose, and their responses to treatment are observed. If no DLTs are observed, another group of three patients receive the next higher dose, and their responses are observed. If one DLT is observed, another three patients are dosed at the same dose to provide reassurance that the DLT was dose-related. If so, the dose is reduced is subsequent patients to refine the definition of the MTD. Once at least one DLT is observed in a group of patients and confirmed in a second cohort of three patients, the dose is reduced in subsequent patients to identify a well-tolerated dose producing DLTs in approximately the desired percentage of patients. See Table 7.6. Note that one patient did not report for dosing in the third dose group, so only two patients were dosed.

Table 7.6: Example 7.2.3: Dose Limiting Toxicity Data from a First-time-in-humans Trial

Subject	Dose(mg)	DLT
1	1	0
101	1	0
2	1	0
3	2	0
102	2	0
4	2	0
103	4	0
5	4	0
6	6	0

DLT=1 DLT Observed
DLT=0 DLT not Observed		

Table 7.6: Example 7.2.3: Dose Limiting Toxicity Data from a First-time-in-humans Trial

Subject	Dose(mg)	DLT
104	6	0
7	6	0
8	8	0
105	8	1
106	8	0
107	8	0
9	8	0
10	8	0
108	10	0
11	10	0
12	10	0
109	12.5	0
110	12.5	0
13	12.5	0
111	16	0
112	16	0
14	16	0
15	16	0
113	21	1
114	21	1
16	21	1
17	18	0
18	18	0
19	18	0
115	18	0
20	18	0
116	18	0
21	18	1
22	18	0
23	18	1
117	18	0
118	18	0
24	18	1
25	18	1
26	18	0
119	18	0

DLT=1 DLT Observed
DLT=0 DLT not Observed

DLTs are denoted as occurring (1) or not occurring (0) for each individual patient in Table 7.6. Note that these studies are not placebo controlled, and are generally conducted open-label or with only the patients blinded to treatment. Such DLT data is considered as 'Binomial' data (denoting a 0 or 1 response), and the proportion of DLTs as a function of dose may be modelled using a technique known as logistic regression.

To do so, the proportion (P) is defined such that

$$P = \frac{1}{1 + e^{-(\alpha+\beta(ld))}}$$

where β is the slope of a regression of $\ln(P/1-P) = L$ (known as a logit-transformation) on logDose such that $L = \alpha + \beta(ld)$. The parameter α is the intercept at $ld = 0$.

Analysis is straightforward using `proc genmod` in SAS as follows (see code below). One calls the data set (specifying in a `DESCENDING` statement that SAS should model the probability that DLT is 1) and instructs SAS to model the DLTs as a function of logDose. The statement `dist=b` informs SAS that DLT is a binomial endpoint, and `link=logit` specifies that a logit transformation should be used.

First-time-in-humans DLT SAS `proc genmod` *Analysis Code Example 7.2.3*

```
proc genmod data=dlt1 DESCENDING;
    model dlt=lndose/dist=b
    link=logit cl alpha=0.1;
    run;
```

SAS output (not listed) yielded an estimate of -10.5083 for α and 3.3846 for β for Example 7.2.3. This yields the dose-response curve for the proportion of DLTs of Figure 7.3.

We can see that the ED_X is approximately

$$e^{\frac{\ln(X/1-X)-\hat{\alpha}}{\hat{\beta}}}$$

For example, the estimated ED_{30} is 17.4 mg in this analysis.

Note that variation is not taken into account (though it could be) in the calculation of the ED_X. A simple means to do so is to bootstrap the data set (see Chapter 5), and derive the ED_X in each bootstrapped data set. The 5th and 95th quartiles of the bootstrapped data sets for ED_X serve as a 90% confidence interval for our estimate of ED_X, in this case the estimated confidence interval from 1000 bootstraps was 14.1 to 25.4 mg. SAS code to perform this analysis is provided on the website accompanying this book.

Similar procedures my be used to model adverse events in cross-over

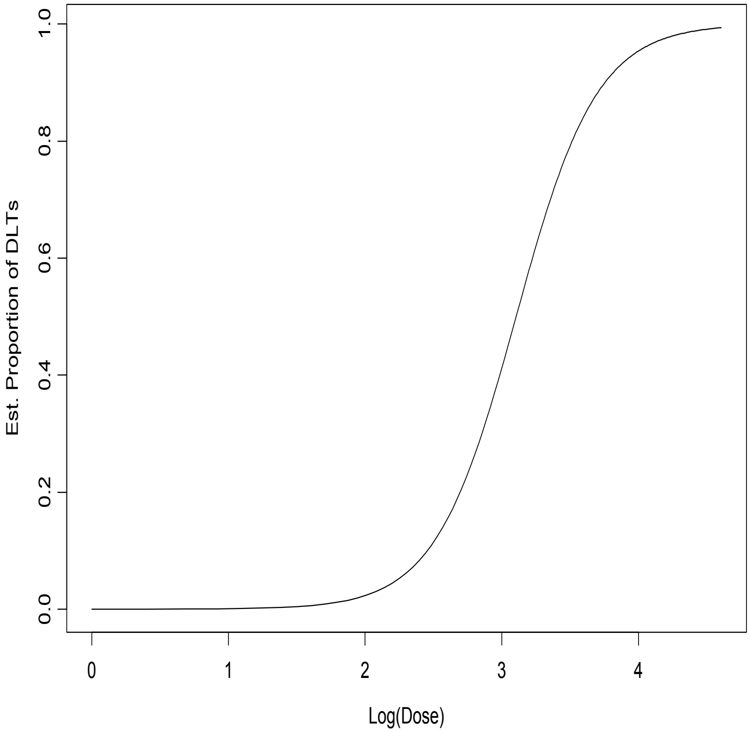

Figure 7.3 *Estimated Proportion of DLTs versus logDose from Example 7.2.3*

trials. See Chapter 6 of [237] for additional details on such techniques. However, given the relative infrequency of AEs in normal healthy volunteer FTiH studies, we do not discuss such application further here.

Intuitively, the use of interactive modelling techniques would seem to add value for such studies. Such techniques utilize data as they are collected, and the models described above, to provide clinicians with an assessment of the safety profile for their choice of future doses. Several techniques have been developed (see Section 7.1) but are infrequently utilised in FTiH studies as experience with them is limited (Chapter 1 [37]). An overview of techniques to aid in decision making in this setting may be found in [456]. See the Technical Appendix for code to perform interactive assessments of PK data in FTiH studies. Those using such interactive techniques are cautioned that 'All models are wrong, but some are useful.' [40] and should note that the use of such techniques supplements, **but in no way should substitute for**, clinical conduct, experience, and expertise. Choice of dose is ultimately a clinical responsibility.

At the end of the FTiH study, the single dose MTD [225] should have been defined. This MTD will possibly be based on observed nonserious AEs, but most likely will be based upon on observed human exposure levels relative to the NOAEL defined in preclinical studies. These studies should definitely provide data to reinforce ideas on the properties of the drug's pharmacokinetics with dose in relation to the NOAEL. In some cases, evidence of pharmacodynamic activity will also be observed, and we will consider methods for modelling of such data in Chapter 9.

Note that the MTD, once defined in this study, is not a constant. As knowledge about the drug accumulates while drug development progresses, it can (and most likely will) change, as can the NOAEL. We now turn to the next study which typically occurs in Phase I.

7.3 Sub-chronic Dosing Studies

Following the FTiH study, a sub-chronic (sometimes referred to as a 'repeat'-dosing) study is performed. The main intent of this trial is to confirm that the MTD defined in the FTiH trial holds true upon repeated administration. In this study again, pharmacokinetic and safety data are most of interest though pharmacodynamic data may be collected if appropriate. Level of blinding (open-label, single-blind, etc.) and choice of population are generally the same as in the FTiH trial. It is unusual for such trials to involve the dosing of patients with the disease for which the treatment is intended. Most often, normal healthy volunteers are dosed for this purpose as in the FTiH trial.

Eligible subjects are randomized to receive either placebo or a dose of drug up to the MTD defined in the FTiH trial. Each dose is administered to 9 to 12 subjects in a cross-over fashion. In the first period, a single dose is given, and pharmacokinetic measurements are collected out to at least five half-lives. Following this, in the second period, subjects receive the same dose at regular, repeated intervals for at least five half-lives, and pharmacokinetic measurements are taken following the last dose over the sampling interval. Following an evaluation of the data collected in the first cohort of 9 to 12 subjects (see Table 7.7), the next highest dose is administered for the next cohort up to the MTD identified in the FTiH trial. The placebo treatment is included to provide a control group for the purposes of safety assessment comparisons, and we will consider an example later where effects were observed in liver function.

The first order of analysis is to assess whether *clearance* is the same after the single dose and after repeated doses. The dose of drug divided by AUC defines a pharmacokinetic parameter known as Clearance (Cl).

Table 7.7 *Schematic Plan of a Sub-chronic Dosing Cross-over Study*

Subject	Period 1	Period 2
	Cohort 1	
1	D1	RD1
2	P	RP
3	D1	RD1
4	D1	RD1
5	P	RP
6	D1	RD1
7	D1	RD1
8	P	RP
9	D1	D1
	Cohort 2	
11	MTD	RMTD
12	MTD	RMTD
13	P	RP
14	MTD	RMTD
15	P	RP
16	MTD	RMTD
17	MTD	RMTD
18	P	RP
19	MTD	RMTD

P=Single Dose of Placebo
RP=Repeated Doses of Placebo
D1=Single Dose of Well-Tolerated Dose from FTiH
RD1=Repeated Doses to Steady State
MTD=Single Dose of MTD from FTiH
RMTD=Repeated Doses of MTD from FTiH to Steady State

More precisely, for an orally dosed drug,

$$Cl_s = \frac{F(dose)}{AUC(0-\infty)},$$

following a single dose of drug (subscript s), denoting the volume of blood cleared of drug in a unit of time for a single dose. The parameter

F is absolute bioavailability (discussed in Chapter 10). When such a drug is dosed repeatedly to steady state, the pharmacokinetic collections on the final dosing day provide an estimate for

$$Cl_{ss} = \frac{F(dose)}{AUC(0-\tau)},$$

where τ is the frequency of dosing (24 h if dosed once a day, 12 h if dosed twice a day) and subscript ss denotes steady state. Steady state concentrations are achieved when the rate of drug being eliminated from the body equals the amount of drug dosed (e.g., dose/hour). In general, this occurs when the drug is dosed repeatedly for at least five half-lives at regular intervals (see Chapters 1 and 2 for a definition of pharmacokinetic half-life).

If $Cl_s = CL_{ss}$ or equivalently in this setting $AUC(0-\tau) = AUC(0-\infty)$ for all doses, then the drug has the property of stationarity of clearance. This is desirable as it makes the drug very easy to dose if the drug has this property. All else being equal, one can be started on a dose estimated to achieve safe and effective drug concentrations, and these concentrations may be maintained by simply taking the same dose at regular intervals. In contrast if $AUC(0-\tau)$ is larger than $AUC(0-\infty)$ then the starting dose might need to be reduced to maintain safe concentrations relative to the NOAEL over time when dosing repeatedly.

Our first example (7.3.1 in Table 7.8) consists of AUC and Cmax data from a sub-chronic dosing study where nine subjects received a dose of either 5, 10, or 20 mg in the first period (accompanying placebo treated subjects are omitted from this discussion as they did not contribute pharmacokinetic data). In the second period, these subjects received the same dose of drug once a day for seven days. On day seven, pharmacokinetic measurements were taken just prior to last the last dose and over the next 24 hours.

Table 7.8: Example 7.3.1: AUC and Cmax Data from a Sub-chronic Dosing Cross-over Study Design

Subject	Dose	$AUC(0-inf)$ S	$AUC(0-\tau)$ SS	Cmax S	Cmax SS
47	5	2.81	5.11	0.267	0.423
48	5	6.31	8.13	0.415	0.620
49	5	7.26	8.01	0.468	0.627
50	5	3.60	6.67	0.410	0.480
52	5	6.82	7.38	0.356	0.591
53	5	1.76	5.17	0.225	0.390
S=Single Dose, SS=Steady State					

Table 7.8: Example 7.3.1: AUC and Cmax Data from a Sub-chronic Dosing Cross-over Study Design

Subject	Dose	$AUC(0-inf)$	$AUC(0-\tau)$	Cmax	Cmax
		S	SS	S	SS
54	5	6.11	8.16	0.471	0.569
55	5	6.09	6.23	0.409	0.483
57	5	2.10	3.36	0.316	0.316
60	10	9.33	11.22	0.820	0.962
61	10	7.31	8.21	0.624	0.723
62	10	9.57	20.85	0.625	1.861
64	10	15.62	16.48	0.798	1.169
65	10	5.56	6.79	0.493	0.574
66	10	11.81	18.08	0.576	1.303
69	10	7.23	10.51	0.723	0.883
71	10	8.35	13.97	0.583	1.056
72	10	5.70	13.80	0.585	1.157
95	20	12.92	30.35	1.514	2.220
99	20	26.05	53.11	2.009	3.902
102	20	23.12	38.61	1.562	2.517
104	20	12.32	29.33	1.002	2.219
105	20	16.35	26.20	1.181	1.844
106	20	20.21	29.47	1.360	1.893
107	20	13.53	27.55	0.970	1.965
108	20	7.70	19.97	0.744	1.447
110	20	14.22	35.91	0.988	2.322
S=Single Dose, SS=Steady State					

For this type of design, the power model is:

$$y_{jk} = (\alpha + \xi_k) + \beta_1(ld) + \phi_j + \beta_2(ld(\phi_j)) + \varepsilon_{jk},$$

where β_1 is the slope parameter of interest regressed on logDose (parameter ld), α and ξ_k are defined as in Section 7.2, ϕ_j denotes the day being studied (j denotes repeat or single dose), β_2 is the slope regressed on logDose on each study day (to account for potential heterogeneity between days within-subjects), and ε_{jk} denotes within-subject error as described in Chapter 3 for each logAUC or logCmax (y_{jk}). If repeat dosing does not impact logAUC or logCmax, then ϕ and β_2 should be zero. Under those circumstances, the model reduces to the same form used in Section 7.2.

Implementation in SAS is straightforward. `proc mixed` is called, and subject and day are specified as classifications. Each endpoint (logAUC or logCmax) is then modelled as a function of logDose, day, and the interaction between logDose and day. Subject is specified as the random-intercept as was done previously using the `random` statement, and desired estimated for the mean effect at each day are output using the `lsmeans` statement. Note that an `at` statement is included in each `lsmeans` statement to instruct SAS to derive estimates at the appropriate choices of logDose (corresponding to doses of 5, 10, and 20) and compare these between days.

Sub-chronic Pharmacokinetic Data Analysis 7.3.1 - SAS `proc mixed` *code:*

```
proc mixed data=pk method=reml;
    class subject day;
    model lnauc=lndose day lndose*day
    /ddfm=kenwardroger s cl alpha=0.1;
    random intercept/subject=subject;
    lsmeans day/at lndose=1.6094 diff cl alpha=0.1;
    lsmeans day/at lndose=2.3026 diff cl alpha=0.1;
    lsmeans day/at lndose=2.9957 diff cl alpha=0.1;
    run;

proc mixed data=pk method=reml;
    class subject day;
    model lncmax=lndose day lndose*day
    /ddfm=kenwardroger s cl alpha=0.1;
    random intercept/subject=subject;
    lsmeans day/at lndose=1.6094 diff cl alpha=0.1;
    lsmeans day/at lndose=2.3026 diff cl alpha=0.1;
    lsmeans day/at lndose=2.9957 diff cl alpha=0.1;
    run;
```

SAS output (not shown) estimates of parameters may be found in Table 7.9. The parameters $\hat{\alpha}$ are the common intercept (response at logDose of zero following repeated dosing), and $\hat{\phi}$ is adjustment to this response following a single dose. The sum of $\hat{\alpha}+\hat{\phi}$ should approximately coincide with the intercept obtained from the FTiH trial, all else being equal (i.e., if formulation or other factors like the pharmacokinetic assay have not changed between trials). The MTD relative to the NOAEL for repeat dosing may be derived as:

$$e^{\frac{\ln NOAEL-\hat{\alpha}}{\hat{\beta}_1+\hat{\beta}_2}}$$

in this design. Confidence intervals for the MTD may again be derived using the bootstrap.

The assessment of stationarity of clearance is accomplished using the findings of the `lsmeans` statements, and relevant outputs may be found

Table 7.9 *Parameter Estimates from Example 7.3.1*

Endpoint	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\phi}$	$\hat{\beta}_2$	$\hat{\sigma}^2_W$
AUC	-0.035	0.93	-0.035	0.23	0.04
Cmax	-2.43	0.87	-0.02	0.21	0.03

in Table 7.10 for logAUC. It was observed that clearance was clearly not stationary for this drug as $AUC(0-\tau)$ was significantly larger than $AUC(0-\infty)$, and accumulation appears to increase with increasing dose. Results on the natural scale may be obtained by exponentiating the below findings. The assessment for Cmax is left as an exercise for interested readers.

Table 7.10 *Stationarity of Clearance Assessment from Example 7.3.1*

Dose	logDose	logAUC(0-τ)-logAUC(0-∞)	90% CI
5	1.61	0.34	(0.19, 0.49)
10	2.30	0.50	(0.40, 0.59)
20	3.00	0.66	(0.51, 0.81)

In our second example, we turn to modelling of the properties of the pharmacokinetic concentration versus time curve. In this study, modelling of this curve generally initiates as the data are rich compared to that collected in later patient studies (where sparse sampling schemes may be employed, see [128]). To clarify, subsequent studies in patients may not be able to employ an extensive pharmacokinetic data collection, as done in Phase I, as it is not convenient to keep patients in-clinic for the lengthy period needed to collect a full pharmacokinetic profile. The profile is modelled in the sub-chronic dosing studies so that a profile can be simulate for a patient population when sparse collections are obtained in subsequent studies.

In the sub-chronic dosing study, each subject receiving an active dose of drug (not placebo) should contribute a drug concentration in plasma versus time profile, as shown in Table 7.11 for Subject 47. Additional data from this study may be found in `conc.sd2` on the website accompanying this book.

We will choose here to utilize SAS for the nonlinear mixed effect mod-

Table 7.11 *Pharmacokinetic Concentration Data from Subject 47 of* `conc.sd2` *following a Single Dose of 5 mg*

Subject	Dose	Time	Conc. ng/mL
47	5	0	.
47	5	0.25	0.117
47	5	0.5	0.221
47	5	0.75	0.266
47	5	1	0.267
47	5	1.5	0.232
47	5	2	0.19
47	5	4	0.178
47	5	6	0.125
47	5	8	0.138
47	5	10	0.145
47	5	12	0.126
47	5	16	0.079
47	5	24	0.051
47	5	36	.
47	5	48	.
47	5	72	.
47	5	96	.

elling of such data; however, several other statistical packages are readily available (SPLUS, NONMEM, WINNONLIN, PKBUGS, etc., [365]) and may be used for this purpose. The models employed are non-linear (as obviously the concentration over time is not linear) and mixed effect in that each subject has an individual profile. Readers interested in more details should see [468] and [435].

For this type of design, we will model the available pharmacokinetic data using what is known as a one-compartment (Chapter 10, [12]) non-linear mixed effect model for the purposes of illustration based on the SAS procedure described in [368] for `proc nlmixed`. Interested readers may use the data in `conc.sd2` on the website accompanying this book to evaluate alternative models. This model assumes that drug is absorbed into the body according to rate k_{ai} (where i denotes subject) and is eliminated from the body according to rate k_{ei}. Concentration c_{it} at time t

for subject i is modelled as follows:

$$c_{it} = (e^{-k_{ei}t} - e^{-k_{ai}t})\frac{k_{ei}k_{ai}(Dose)}{Cl_i(k_{ai} - k_{ei})} + \varepsilon_{it},$$

where ε_{it} represents within-subject residual-error, Cl_i is the clearance for subject i assumed to be of the form $e^{\beta_1+b_{1i}}$ with β_1 being an unknown constant adjusted for each subject as appropriate to b_{1i}. Similarly, k_{ai} is considered to be a function of the form $e^{\beta_2+b_{2i}}$, and k_{ei} is considered to be a function of the form $e^{\beta_3+b_{3i}}$. The parameters b_{1i}, b_{2i}, and b_{3i} are considered to be independent random normal variables with null mean and some nonzero variance in similar fashion to the REML methods described for bioequivalence in Chapter 5.

Implementation in SAS is straightforward. First the data should be sorted by subject to accommodate SAS requirements. The SAS procedure `proc nlmixed` is then called, and following the specification of starting values, the equation described above is specified. Note that here we have assumed concentration is normally distributed. It may be more appropriate to model concentration as log-normally distributed, and this can be accomplished by a log-transformation in a data step. Similarly, instead of modelling concentration as a function of dose, logDose may be more appropriate.

Nonlinear Mixed Effect Pharmacokinetic Data Analysis of Phase 1 Concentration Data in `conc.sd2` *- SAS* `proc nlmixed` *code:*

```
proc sort data=my.conc;
    by subject dose time;run;

proc nlmixed data=my.conc;
    parms beta1=0.4 beta2=1.5 beta3=-2 s2b1=0.04
    s2b2=0.02 s2b3=0.01 s2=0.25;
    cl = exp(beta1+b1);
    ka = exp(beta2+b2);
    ke = exp(beta3+b3);
    pred=dose*ke*ka*(exp(-ke*time)-exp(-ka*time))/
    (cl*(ka-ke));
    model conc ~ normal(pred,s2);
    random b1 b2 b3 ~ normal([0,0,0],[s2b1,0,
    s2b2,0,0,s2b3]) subject=subject;
    predict pred out=pred;
    run;
```

In this code, s2b1, s2b2, and s2b3 are the variances associated with b_{i1}, b_{i2}, and b_{i3}, respectively. The parameter s2 is the estimate of within-subject variance. Estimated parameters may be found in Table 7.12, and a plot of the estimated concentrations for each dose versus time may be found in Figure 7.4.

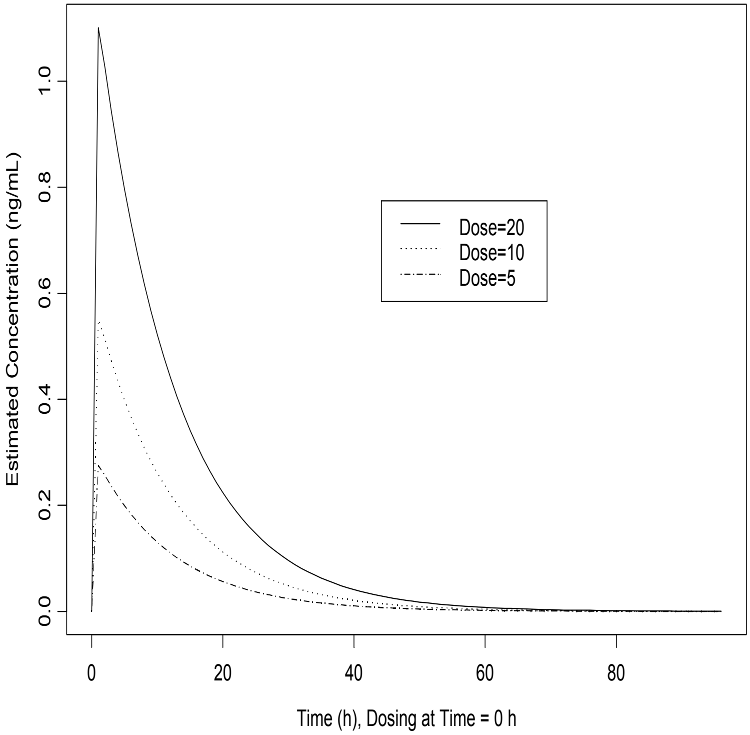

Figure 7.4 *Estimated Concentration versus Time (h) Profile from Phase 1 Concentration Data in* `conc.sd2`

Predicted concentrations from the model are output to a data set `pred` using the statement `predict pred out=pred;` in the above code. These values may be used to construct residual plots for each subject and across subjects to assess model fit using the following SAS code. Some evidence of poor model fit is evident at low concentrations; however overall, the model appears to provide an adequate description of the data.

Table 7.12 *Estimated PK Model Parameters from Phase 1 Concentration Data in* `conc.sd2`

Parameter	Estimate	95% CI
β_1	0.35	0.23,0.47
β_2	1.46	1.30,1.63
β_3	-2.47	-2.58,-2.36
s2b1	0.04	0.01,0.08
s2b2	0.03	-0.02,0.09
s2b3	0.01	-0.01,0.02
s2	0.011	0.009,0.013

Nonlinear PK Analysis Model Diagnostic Code:

```
proc sort data=pred;
    by subject dose time;run;
data pred;set pred;
    st_resid=(conc-Pred)/StdErrPred;
    run;
proc rank data=pred normal=blom out=nscore;
    var st_resid;
    ranks nscore;
data nscore;
  set nscore;
  label nscore="Normal Score";
  label stres="Residual";
  label pred="Predicted Value";
    run;
proc plot vpercent=50 data=nscore;
    plot st_resid*pred/vref=0;
    plot st_resid*nscore;
    run;
```

In subsequent studies, when limited concentration data are collected from patients at a given time on a given dose, these data can be used with the model findings above to simulate a population pharmacokinetic profile. This can then be used to assess the exposure levels in that patient population relative the NOAEL, and we will discuss how such assessments may be done in Chapter 10. Similar models are used to characterize the concentrations after repeat dosing. Clearance is differentiated between single and repeat dosing as appropriate to the findings of the stationarity of clearance assessment.

We now consider findings of ALT elevation which were observed in a repeat dose trial. ALT elevations are potentially indicative of liver injury, and were monitored each day in this study. Such elevations can occur spontaneously and unpredictably, in response to strenuous exercise, for instance. Of concern here, however, was that these elevations were presumed to be drug induced. Although the ALT returned to baseline upon cessation of treatment (data not shown), it was of interest to model the behavior of ALT with dose over time to provide clinical with a means of designing a monitoring plan in subsequent studies. For this assessment, we will treat ALT as being log-normally distributed and model it as a function of logDose.

Data from Subject 4 (who received 50 mg) may be found in Table 7.13. The data for the remaining subjects may be found in the data set `liver.sd2` on the website accompanying this book. For this subject we see little indication of a response to drug treatment until day 5 whereupon the ALT begins to increase.

Table 7.13 *ALT Data from Subject 4 of* `liver.sd2`

Subject	Period	Dose	Day	ALT
4	2	50	1	13
			2	13
			3	16
			4	15
			5	18
			6	24
			7	25
			8	29
			9	34
			10	36
			11	34
			12	33
			13	45
			14	43

For this type of design, the power model for ALT is an extension of the model used for pharmacokinetic data earlier in this section:

$$y_{jk} = \alpha + \phi_j + \beta_1(ld) + \beta_2(ld(\phi_j)) + \varepsilon_{jk},$$

where ϕ_j denotes the day being studied (j denotes days 1 to 14) for each logALT (y_{jk}). If dosing does not impact logALT, then β_1 and β_2 should be zero. We presume that ALT responses from day to day within a subject are related to each other, with the degree of correlation decreasing with increasing time between days, and will partition this aspect of variance associated with ϕ_j from the within-subject variation ε_{jk} in our model.

Here, `proc mixed` is called, and subject and day are specified as class variables. The endpoint of interest (logALT) is then modelled as a function of logDose, day, and the interaction between logDose and day. The correlation between days is partitioned from the within-subject variance using a `repeated` statement specifying that the correlation occurs within each subject. The desired estimates for the mean effect at each day are output using the `lsmeans` statement. Note that an `at` statement is again included in each `lsmeans` statement to instruct SAS to derive estimates at the appropriate choices of logDose (corresponding to doses of approximately zero to 3000).

Sub-chronic ALT Data Analysis of `liver.sd2` - *SAS* `proc mixed` *code:*

```
proc mixed data=liver;
    class subject day;
    model lnalt=day lndose day*lndose
    /DDFM=KENWARDROGER S outp=out;
    repeated day/type=AR(1) subject=subject;
    lsmeans day/at lndose=-11.5129 CL alpha=0.01;
    lsmeans day/at lndose=3.91 CL alpha=0.01;
    lsmeans day/at lndose=4.61 CL alpha=0.01;
    lsmeans day/at lndose=5.01 CL alpha=0.01;
    lsmeans day/at lndose=5.52 CL alpha=0.01;
    lsmeans day/at lndose=6.21 CL alpha=0.01;
    lsmeans day/at lndose=6.62 CL alpha=0.01;
    lsmeans day/at lndose=6.91 CL alpha=0.01;
    lsmeans day/at lndose=7.60 CL alpha=0.01;
    lsmeans day/at lndose=8.01 CL alpha=0.01;
    ods output LSMeans=my.means1;
    run;
```

In this data set, for this population (recall these are normal healthy volunteers), statistically significant logDose related ($p = 0.0415$) increases in ln-ALT were observed, and these changes increased with increasing dose ($p = 0.0024$). The estimates of ALT elevation for the 50 mg (logDose of 3.91) and the 3000 mg dose (logDose of 8.01) are presented in the following table, exponentiated back to the original scale.

Table 7.14: Estimated ALT Data (based on `liver.sd2`) from the Sub-chronic Dosing Study Design

Dose	Day	Est. ALT	90% CI
50	1	15.5	12.9,18.5
50	2	14.9	12.5,17.9
50	3	15.2	12.7,18.2
50	4	17.1	14.3,20.5
50	5	20.9	17.5,25.1
50	6	25.4	21.2,30.5
50	7	28.3	23.6,33.9
50	8	29.4	24.5,35.2
50	9	30.1	25.1,36.1
50	10	31.1	26.0,37.3
50	11	31.1	26.0,37.3
50	12	29.5	24.6,35.3
50	13	30.6	25.5,36.6
50	14	31.5	26.2,37.7
3000	1	15.4	12.3,19.2
3000	2	14.9	11.9,18.6
3000	3	15.2	12.1,18.9
3000	4	17.3	13.9,21.7
3000	5	22.1	17.7,27.7
3000	6	27.9	22.3,34.8
3000	7	31.8	25.5,39.8
3000	8	33.3	26.6,41.6
3000	9	34.4	27.5,43.1
3000	10	35.4	28.3,44.2
3000	11	34.9	27.9,43.7
3000	12	32.6	26.1,40.8
3000	13	33.5	26.8,41.9
3000	14	34.5	27.5,43.2
Upper Limit of Normal ALT=34			

For the 50 mg dose, we see ALT elevations beginning on day 4 and continuing throughout the dosing interval. Potentially hazardous elevations may be expected seven days after beginning dosing (when the 90% upper bound crosses the upper limit of normal). ALT elevations were slightly greater as dose was increased to 3000 mg and potentially hazardous ALT elevations were encountered a day sooner.

Models similar to these may be used to test for proof of safety (see Chapter 8) and to model the behavior of pharmacodynamic effects (see Chapter 9). We now turn to another safety topic.

7.4 Food-Effect Assessment and DDIs

Following the studies described in the previous two sections, the maximum tolerated dose should have been identified when a single dose of drug has been given and when a dose of drug is given repeatedly. By that point, drug developers should have a good handle on what the body does to the drug in isolation.

Note that what has not been done at this point is as important as what has been learned. Drug development at this stage should have confirmed that the potential for hazard when taking the drug is low when given at certain doses over a period of limited duration. If a potential hazard with dose has been identified, it may be necessary to explicitly study the drug to provide 'proof of safety' under a variety of potential clinical uses - see Chapter 8 for one such example.

Other preclinical and clinical studies later in development will be needed if the drug is to be given chronically for longer intervals. Additionally, the behavior of the drug in people with disease and different ethnicity (Chapter 10) has not yet been established.

However, no one actually takes a drug in isolation. Patients are expected to take the drug with food on occasion and may be expected to take it while taking other agents (whether or not the label precludes such [426]). In this context, alcohol is an agent; over-the-counter vitamins and pharmaceuticals are other examples of agents, etc. How the body handles the drug when coadministered under such circumstances is the subject of this section.

As we know (Chapter 2), when a drug is taken it undergoes absorption, distribution, and metabolism and is eventually eliminated from the body (ADME). Dosing a drug with food may impact how the drug is absorbed. Dosing of a drug with other agents can impact distribution and, more frequently, metabolism. This can slow down or speed up elimination of the drug substance from the body. If elimination is decreased, exposure to drug may increase to the point where it is not well tolerated. Alternatively, if elimination is enhanced, the dose of drug may not be sufficient to cause an efficacious response.

Lack of a meaningful pharmacokinetic difference when a drug product is administered with and without food or with and without a concomitantly administered agent or medication can often be assessed using the results of small cross-over studies and applying a TOST approach [416]. Rate of bioavailability as measured by Cmax is held, under this approach, to be a surrogate marker for safety for drugs in the marketplace. Comparable or decreased mean Cmax following administration with or without food or a concomitantly administered medication are indicative of similar safety hazards to that when dosed alone. Increases in mean Cmax are potentially suggestive of a less acceptable safety profile for the drug under study. Similarly, comparable mean AUC following administration with or without food or a concomitantly administered medication are indicative of safety and efficacy in that condition. The magnitude of decrease or increase in exposure can be used to adjust the dosing strategy for the drug product under study.

As with bioequivalence, pharmacokinetics serve as a tool for assessing safety in this context. Such a assessment limits the potential for hazard established in the First-time-in-humans and sub-chronic dosing studies, but does not eliminate it entirely.

We first consider an example of a cross-over study assessing the potential for dosing with a meal to impact exposure (food effect). This is followed by two examples of drug interaction trials.

Dosing of a drug product with a meal can change absorption of the drug substance by [132]:

1. Delaying gastric emptying,
2. Stimulating bile flow,
3. Changing gastrointestinal PH,
4. Increasing splanchnic blood flow,
5. Changing luminal metabolism,
6. Causing physical or chemical interactions with the formulation or drug substance

The effect of food on absorption is typically studied using an open-label, randomized, 2 × 2 cross-over trial in normal healthy volunteers. See Chapter 3 and [132] for details. Subjects (normal healthy volunteers) are randomized to receive one of two sequences of treatment regimens. Subjects receive a dose of drug following an overnight fast, are washed out for five half-lives, and then receive the same dose of drug following a meal, or vice-versa.

Note the change in terminology in this section to *regimen* instead of formulation. In a food effect study, the formulation is the same, only the conditions of dosing (with or without a meal) are changed. The use of the descriptor regimen denotes that the dose of drug under study is the same, but study conditions are altered to study the ADME properties. In Example 7.4.1 (below), regimens A and B denote dosing without (regimen A) and with (regimen B) a meal. As with bioequivalence testing, absence of a food effect is concluded if the 90% confidence intervals for AUC and Cmax $\mu_B - \mu_A$ fall within the standard bioequivalence acceptance limits of $-\ln 1.25, \ln 1.25$ [132].

We now turn to an example of such testing for food effect. In this trial (Example 7.4.1), 20 normal healthy volunteers were randomly assigned to sequences AB and BA, and AUC and Cmax were measured following dosing in each period.

Table 7.15: Example 7.4.1: AUC and Cmax Data from a 2×2 Food Effect Cross-over Study Design

Subject	Seq	AUC A	AUC B	Cmax A	Cmax B
1	AB	5836	8215	1953	1869
2	BA	9196	9895	1769	2446
3	AB	7809	7222	3409	1501
4	BA	6443	18864	1916	4232
5	BA	5875	5911	1884	2087
6	AB	9937	6186	2807	1743
7	BA	10275	9135	2532	2736
8	AB	4798	6211	1912	1541
9	BA	8940	9810	1939	2216
10	AB	10739	14734	1908	3645
11	AB	10549	10937	4042	2120
12	BA	8374	10853	3702	2001
13	BA	16510	13205	3411	2840
14	AB	7534	5648	2119	1684
15	AB	9473	13407	4194	3074
16	BA	5118	9399	2294	1538
17	AB	4686	7504	1487	1839
18	BA	6122	11027	1857	2063
19	AB	14059	15765	3142	3120
20	BA	6841	8104	1883	1954
A=Fasted Dose, B=Fed Dose					

Data were analyzed using the procedures of Chapter 3 based on the following `proc mixed` code.

Food Effect Example 7.4.1 - SAS `proc mixed` *code:*

```
proc mixed data=pk_food;
    class sequence subject period regimen;
    model logauc=sequence period regimen/
    ddfm=kenwardroger;
    random subject(sequence);
    lsmeans regimen/pdiff cl alpha=0.1;
    estimate ’Food Effect for logAUC’ regimen -1 1;
    run;

proc mixed data=pk_food;
    class sequence subject period regimen;
    model logcmax=sequence period regimen/
    ddfm=kenwardroger;
    random subject(sequence);
    lsmeans regimen/pdiff cl alpha=0.1;
    estimate ’Food Effect for logCmax’ regimen -1 1;
    run;

proc mixed data=pk_food;
    class sequence subject period regimen;
    model tmax=sequence period regimen/
    ddfm=kenwardroger;
    random subject(sequence);
    lsmeans regimen/pdiff cl alpha=0.1;
    estimate ’Food Effect for Tmax’ regimen -1 1;
    run;
```

Dosing with food significantly ($p = 0.0363$) increased the extent of exposure (AUC) to this drug product by approximately 20% with an estimate of food effect ($\mu_B - \mu_A$) of 0.1788 (90% confidence interval 0.0417, 0.3158) on the log-scale. Although rate of exposure (Cmax) was not significantly changed ($p = 0.4142$), lack of food effect could not be concluded as the estimate of food effect was -0.0758 (90% confidence interval -0.2330, 0.0814) on the log-scale. Tmax was significantly prolonged following dosing with a meal (data may be found on the website accompanying this book) with food effect estimated to be 1.7h (90% confidence interval 1.28h, 2.11h).

From these data, it is possible to conclude that dosing with food affects the absorption of this drug product, increasing the overall exposure to drug (AUC) and delaying its maximal concentration. These changes do not likely present a hazard to patients using the drug as Cmax was not increased following a meal, and the increase in AUC was not deemed clinically relevant (requiring a change in dose to correct).

We now turn to the statistical assessment of drug interactions. Drugs can interact with each other in a number of ways involving the ADME properties ([37] Chapter 2, [12] Chapter 14). As with food effects, absorption may be impacted; however, the most common interaction relates to how the liver metabolizes the drug substances. Metabolic inhibition denotes that one drug prevents the metabolism of the other, usually resulting in increased exposure to the substance. Alternatively, drugs may have no effect on each other or a drug might induce the metabolism of the other indicating that metabolism activity is enhanced in the body likely leading to decreased exposure to drug.

Note that metabolism is only one way that drugs can interact. Other examples include protein binding interactions, transporter interaction, etc. See [37] Chapter 2 and [12] Chapter 14 for more details. In this section, we will discuss the topic of drug interactions focusing on those introduced by the CYP450 liver enzyme system for simplicity; however, the clinical and statistical assessments used are similar for these other interaction types.

The CYP450 (cytochrome P450) enzyme family is responsible for the majority of metabolic drug interactions known to occur [127]. This type of drug metabolism is focused in the body's liver, and the liver uses multiple subfamily enzyme systems to metabolize drug products after they are ingested and as they circulate through the blood. The subfamilies include, in decreasing order of importance and frequency [37]:

1. 3A4,
2. 2C9,
3. 2A6,
4. 2C8, 2E1,
5. 1A2,
6. 2B6,
7. 2D6, 2C19, etc.

Inhibition or induction of drugs metabolized by these systems may result in changed exposure levels, presumably and potentially putting the safety of patients at risk. Clinical studies are used to assess this potential.

In vitro testing [123] may preclude or enhance the need to do such a study. The predictive value of such in vitro testing for drug metabolism by the CYP450 family has become increasingly accurate and reliable in recent years, and generally, clinical drug interaction trials are only conducted when an in vitro system identifies a particular subfamily as being of potential concern. Such concern may arise if the new drug inhibits or induces the metabolism of other drugs by a certain subfamily or if the new drug is itself metabolized by a particular subfamily - the route for which may be inhibited or induced by another product.

To assess the potential changes in exposure, a steady state randomized or non-randomized cross-over design are most often used. In general, subjects are dosed to steady state with one product alone (Regimen A in the following examples), and in the alternative regimen are dosed to steady state with the potential metabolic inhibitor or inducer in tandem (Regimen B). AUC, Cmax, and other pharmacokinetic endpoints are derived at appropriate times following dosing to evaluate the potential changes in exposure [127].

Non-randomized cross-over designs (see Example 7.4.3 below) may be used if washout of the probe drug (i.e., the drug being probed for a potential interaction) is long or if an extended dosing period is necessary to achieve steady state exposure. It should be noted that it is possible to administer several probe drugs at the same time to evaluate multiple pathways of metabolism at once. These are known a 'cocktail' drug-interaction trials. See [398] for a recent example.

Our first drug interaction example is a randomized cross-over study in 20 normal healthy volunteers where a probe drug's metabolism was inhibited when given with a new drug at steady state. The increase in exposure was studied to determine whether coadministration represented a risk to patients using the probe drug. SAS code to analyse such data are the same as that applied in Chapter 3 and may be found below.

Table 7.16: Example 7.4.2: AUC and Cmax Data from a 2 × 2 Drug Interaction Cross-over Study Design for Metabolic Inhibition

Subject	Seq	AUC	AUC	Cmax	Cmax
		A	B	A	B
1	BA	21.9	28.1	2.16	2.27
2	AB	17.9	14.8	1.63	1.39
3	BA	14.8	22.2	1.21	2.38
4	AB	19.4	17.0	1.59	1.64
6	AB	28.2	28.2	2.77	2.84
7	AB	25.3	17.1	1.98	1.84
8	BA	24.0	25.4	1.71	1.90
10	AB	27.8	33.2	2.68	2.57
11	BA	17.0	20.6	1.98	2.49
12	AB	19.3	23.6	2.37	3.29
14	AB	29.9	27.5	2.43	2.22

A=Probe Drug
B=Probe Drug Plus a Metabolic Inhibitor

Table 7.16: Example 7.4.2: AUC and Cmax Data from a 2×2 Drug Interaction Cross-over Study Design for Metabolic Inhibition

Subject	Seq	AUC A	AUC B	Cmax A	Cmax B
15	AB	20.5	22.3	1.92	2.04
16	BA	24.3	29.9	2.26	2.83
17	BA	27.5	32.5	1.92	2.27
18	AB	16.9	17.4	1.66	1.91
19	AB	33.1	39.0	3.39	2.88
20	BA	14.7	22.1	1.63	2.66
21	BA	29.3	43.2	2.46	3.79
22	AB	23.3	31.6	3.06	2.57
23	BA	23.1	24.3	2.66	2.56
A=Probe Drug B=Probe Drug Plus a Metabolic Inhibitor					

Inhibitor Drug Interaction Example 7.4.2 - SAS `proc mixed` *code:*

```
proc mixed data=pk_inhi;
    class sequence subject period regimen;
    model logauc=sequence period regimen/
    ddfm=kenwardroger;
    random subject(sequence);
    lsmeans regimen/pdiff cl alpha=0.1;
    estimate 'DDI Effect for logAUC' regimen -1 1;
    run;

proc mixed data=pk_inhi;
    class sequence subject period regimen;
    model logcmax=sequence period regimen/
    ddfm=kenwardroger;
    random subject(sequence);
    lsmeans regimen/pdiff cl alpha=0.1;
    estimate 'DDI Effect for logCmax' regimen -1 1;
    run;
```

Dosing with the metabolic inhibitor significantly changed AUC and Cmax of the probe drug ($p = 0.0056$ and 0.0094, respectively). Administration with the metabolic inhibitor increased the extent of exposure (AUC) to this drug product by approximately 13% with an estimate of interaction ($\mu_B - \mu_A$) of 0.1254 (90% confidence interval 0.0563, 0.1946) on the log-scale. The maximal concentration (Cmax) was also increased by 13% with effect size of 0.1245 (90% confidence interval 0.0503, 0.1987) on the log-scale. Other data (Tmax, etc.) measured in this study may be found on the website accompanying this book. Interested readers should note that C24 (the concentration of probe drug 24 hours following dosing) and renal clearance (CLR) were significantly altered by combination dosing; however, Tmax was not.

Our second drug interaction example is a non-randomized cross-over study in 20 normal healthy volunteers where a probe drug's metabolism was induced when given with a new drug at steady state. The decrease in exposure was studied to determine whether coadministration represented a risk to patients using the probe drug. SAS code to analyse such data are similar to that applied in Chapter 3 and may be found below. Note, this was a non-randomized cross-over study, so period and sequence effects are confounded with regimen (and were not fitted in the model). This type of design is acceptable [127] when period effects can be expected to be small relative to the effect of regimen.

Table 7.17: Example 7.4.3: AUC and Cmax Data from a Drug Interaction Cross-over Study Design for Metabolic Induction

Subject	Seq	AUC A	AUC B	Cmax A	Cmax B
1	AB	37.73	9.38	3.84	2.75
2	AB	18.22	5.07	2.74	0.97
3	AB	10.30	5.75	1.87	1.98
4	AB	22.11	4.32	4.32	1.15
5	AB	16.31	5.83	3.24	1.15
6	AB	20.47	6.80	3.23	1.32
7	AB	16.02	3.32	1.71	0.72
8	AB	10.73	3.38	1.99	1.07
9	AB	13.93	3.72	1.92	0.97
10	AB	24.32	4.25	2.99	0.59
11	AB	31.67	6.82	3.03	1.01
12	AB	10.97	3.40	2.03	0.48
13	AB	55.49	7.72	4.90	2.20
14	AB	13.65	4.16	1.73	0.65
15	AB	23.97	6.13	3.27	1.78
16	AB	14.07	2.65	2.65	0.50
17	AB	6.51	2.59	1.32	0.91
18	AB	19.60	3.32	3.07	0.56
19	AB	18.80	2.96	2.83	0.66
20	AB	28.25	3.32	3.11	0.69

A=Probe Drug
B=Probe Drug Plus a Metabolic Inducer

Inducer Drug Interaction Example 7.4.3 - SAS `proc mixed` *code:*

```
proc mixed data=pk_indu;
    class subject regimen;
    model logauc=regimen/ddfm=kenwardroger;
    random subject;
    lsmeans regimen/pdiff cl alpha=0.1;
    estimate ’DDI Effect for logAUC’ regimen -1 1;
    run;

proc mixed data=pk_indu;
    class subject regimen;
    model logcmax=regimen/ddfm=kenwardroger;
    random subject;
    lsmeans regimen/pdiff cl alpha=0.1;
    estimate ’DDI Effect for logCmax’ regimen -1 1;
    run;
```

Dosing with the metabolic inducer significantly changed AUC and Cmax of the probe drug ($p < 0.0001$ for both endpoints). Adminstration with the metabolic inducer decreased the extent of exposure (AUC) to this drug product by approximately 75% with an estimate of interaction ($\mu_B - \mu_A$) of -1.4199 (90% confidence interval -1.5686, -1.2713) on the log-scale. The maximal concentration (Cmax) was also decreased by 63% with effect size of -0.9996 (90% confidence interval -1.1883, -0.8109) on the log-scale. Other data (half-life, Tmax, etc.) measured in this study may be found on the website accompanying this book. Interested readers should note that half-life was significantly altered by combination dosing; however, Tmax was not.

While combination of dosing with these products may be presumed to be safe (as exposure was decreased), it may not be desirable. Changes of this magnitude in exposure might lead to the probe drug being inefficacious, and alternative dosing strategies might need to be employed to ensure adequate probe drug is available in the body to succeed in establishing an effective treatment.

The sample size required to have sufficient power for food effect designs and TOST assessment [132] are derived according to the procedures developed in Chapter 3. However, in drug interaction trials, regulatory guidance generally does not call for TOST assessment relative to the traditional bioequivalence acceptance limits of $-\ln 1.25$ to $\ln 1.25$. More commonly, no-effect boundaries are predetermined by means of assessing how much change in exposure would necessitate a change in dose for the probe drug to be safe and efficacious. These limits need not be symmetric, and SAS code is provided below to perform such a derivation of sample size.

Sample Size Code for TOST in DDI Studies:

```
data a;
    * total number of subjects
    (needs to be a multiple of number
    of sequences, seq);
n=20; seq=2;
    * significance level;
a=0.05;
    * variance of difference of two observations
    on the log scale;
    * sigmaW = within-subjects standard deviation;
sigmaW=0.2; s=sqrt(2)*sigmaW;
    * error degrees of freedom for cross-over
    with n subjects in total
    assigned equally to seq sequences;
n2=n-seq;
    * ratio = mu_T/mu_R;
    ratio=1.00;
lal=0.8;
    *lower acceptance limit;
ul=1.25;
    *upper acceptance limit;
    run;

data b; set a;
* calculate power;
    t1=tinv(1-a,n2); t2=-t1;
    nc1=(sqrt(n))*((log(ratio)-log(lal))/s);
    nc2=(sqrt(n))*((log(ratio)-log(ual))/s);
    df=n2;
    prob1=probt(t1,df,nc1);
    prob2=probt(t2,df,nc2);
    answer=prob2-prob1;
    power=answer*100; run;

proc print data=b; run;
```

Lower and upper acceptance limits are not always available from the literature, and even if they are, regulators may not agree with whatever the sponsor defines. Under such circumstances, an estimation approach

[242] can be useful when the magnitude of no-effect boundaries are not known and the main study objective is to provide evidence of what the potential value, or range of values, may be, or when the sample size is in part set by feasibility, and we wish to provide an idea of the precision the trial is likely to provide for the drug interaction effect of interest.

In such cases, the intent is to provide an estimate of the expected width or precision of the plausible range of values as expressed by a confidence interval. This will help satisfy our expectation with regard to acceptability and applicability of study results in the knowledge that, 'The confidence interval can be thought of as the set of true but unknown differences that are statistically compatible with the observed difference.' [164]

Then, as described in Chapter 3, Equation (3.7), a 90% confidence interval for $\mu_T - \mu_R$ is:

$$\hat{\mu}_T - \hat{\mu}_R \pm t_{0.95}(n-2)\sqrt{\frac{2\hat{\sigma}_W^2}{n}},$$

when sample size in each sequence is equal and n is the overall sample size. For the purposes of this discussion, we presume a standard 2 × 2 cross-over is used, but alteration for alternative designs is easily accomplished and is left as an exercise for the interested reader. Consider

$$w_\delta = t_{0.95}(n-2)\sqrt{\frac{2\hat{\sigma}_W^2}{n}}.$$

This function provides a precision estimate for the true mean difference. Goodman [164] notes that use of a method like that proposed above should be exercised with caution as, in a situation where the study design is truly intended to support a test of hypothesis, the approach corresponds to a test using only 50% power when precision is equal to the difference of interest. Similarly, in situations where a TOST equivalence approach is intended, the method presented in this equation corresponds to a two one-sided hypothesis test with 50% power when precision is equal to the equivalence range of interest.

Sample Size Code for Precision in DDI Studies:

```
data a;
    * total number of subjects
    (needs to be a multiple of number
    of sequences, seq);
n=20; seq=2;
    * significance level;
a=0.05;
    * variance of difference of two observations
    on the log scale;
    * sigmaW = within-subjects standard deviation;
sigmaW=0.2; s=sqrt(2)*sigmaW;
    * error degrees of freedom for cross-over
    with n subjects in total
    assigned equally to seq sequences;
n2=n-seq;
    run;

data b; set a;
* calculate precision;
    t=tinv(1-a,n2);
    SE=s/(sqrt(n));
* precision on log-scale;
    w=t*SE;
* precision on natural-scale;
    exp_w=(exp(t*SE)-1)*100;
    run;

proc print data=b; run;
```

In this case, the precision on the natural scale would be calculated as 12%, indicating that the confidence limits will lie about that far from the point estimate for the difference in means. If greater precision is desired, the sample size may be increased, or decreased if lesser precision is needed.

In some cases, such pharmacokinetic safety assessment will not suffice, and a more rigorous assessment of safety may be called for to protect patients using the drug. Under such circumstances, often a specific biomarker is of interest. Such an example - QTc - will be considered in the next chapter.

7.5 Dose-Proportionality

In developing drugs, sponsoring companies spend a great deal of time and energy mapping pharmacokinetic exposure to drug (concentration in blood, AUC, Cmax, etc.) with clinical outcomes relating to safety and efficacy. When working in a clinical setting, physicians do not often have access to pharmacokinetic data from their patients. In practice, therefore, they vary dose in their patients to cause clinical benefit, and limit dosing to ensure undesirable side-effects (e.g., nausea, emesis) do not occur. Consider a situation where one administers a dose sure to be efficacious, but observes an unacceptable side effect (e.g., nausea). Dose-proportionality, the subject of this section, helps one determine which lower dose should next be tried to improve tolerability while still attaining efficacy.

In the previous studies discussed in this chapter, an understanding of the dose to exposure to safety relationship will have been established. One of the things prescribers need to know is how much exposure changes when the dose is changed, so that in changing doses for a given patient, they can balance a change in dose with desirable outcomes (see Chapter 9) and undesirable side-effects.

When one increases the dose of a drug product, this does not necessarily result in a proportional change in exposure. There are physiologic, biologic, and chemical limits to how much drug substance the body will absorb, distribute, metabolize, and excrete. However, over the therapeutic dose range (the maximum effective and tolerated dose less the minimum effective dose), it is important to know that if one, say, doubles the dose, then double the rate and extent of exposure results - and vice versa. [408].

The assessments of rate and extent of exposure in the first-time-in-humans and sub-chronic dosing studies will yield a good practical understanding of the shape of the dose-to-exposure relationship (as described in previous sections). However, assessments of dose-proportionality in the first-time-in-humans study are confounded with period effects. These effects are known to occur in pharmacokinetic studies and may impact inference [376]. Assessments of dose-proportionality in the sub-chronic dosing study are generally underpowered for robust statistical assessment as the study is parallel group. While knowledge gained from these studies, in general, is adequate for clinical development, for approval at regulatory agencies (in preparation for giving the drug to large populations), a more robust study is generally done to confirm that the shape of the dose-to-exposure relationship is well understood.

In some situations, therefore, a confirmatory dose-proportionality study is performed just prior to regulatory filing with the final to-be-marketed formulation. Many different models may be used to examine dose proportionality [408]. This section will focus on the application of the power model (described in Section 7.2) for this assessment. In this setting, we will assume a randomized cross-over design is used to assess dose-proportionality with at least three doses in the therapeutic range being considered. Normal healthy volunteers receive a dose of drug after an overnight fast, with administration of each dose separated by a washout period of at least five half-lives using a Williams square design (see Chapter 4).

For this type of design, the model is:

$$y_{ijk} = \alpha + \beta(ld) + \pi_j + \gamma_i + \xi_{k(i)} + \varepsilon_{ijk},$$

where α, β, and logDose (ld) are as previously described, π_j and γ_i identify the period j of sequence i, $\xi_{k(i)}$ is the random-intercept accounting for each subject within sequence as their own control, and ε_{ijk} denotes within-subject error as described in Chapter 3 for each log-transformed AUC or Cmax (y_{ijk}).

When one exponentiates both sides of this equation, AUC or $Cmax = c(d^{\beta})$ where c is a value composed of the exponentiated sum of estimates of sequence, subject, and period fixed effects and d is dose. When $\beta = 1$, the drug is dose proportional as AUC or $Cmax = cd$. When one wishes to change the dose, it is easy to predict what AUC or Cmax will result. If $\beta \neq 1$, one can still predict what AUC or Cmax will result from changing the dose, but the calculation is more complex (as the relationship of dose to the exposure endpoint, AUC or Cmax, is nonlinear).

Consider the possible shape of the resulting dose to exposure curves in Figure 7.5.

For $\beta = 1$, a truly dose-proportional relationship is observed. For any unit change in dose, a unit change in AUC results - i.e., doubling the dose results in twice the AUC. If $\beta > 1$, a greater than dose-proportional response is seen (doubling the dose results in a greater than doubling in AUC), and if $\beta < 1$, a less than dose-proportional response in exposure is observed (doubling the dose results in less than a doubling in AUC).

Smith et al. [406] showed that it is obvious to think of dose proportionality as an equivalence problem. This implies that the structure for testing dose proportionality is:

$$H_{01} : \beta \leq 1 - t$$

versus

$$H_{11} : \beta > 1 - t$$

and

$$H_{02} : \beta \geq 1 + t$$

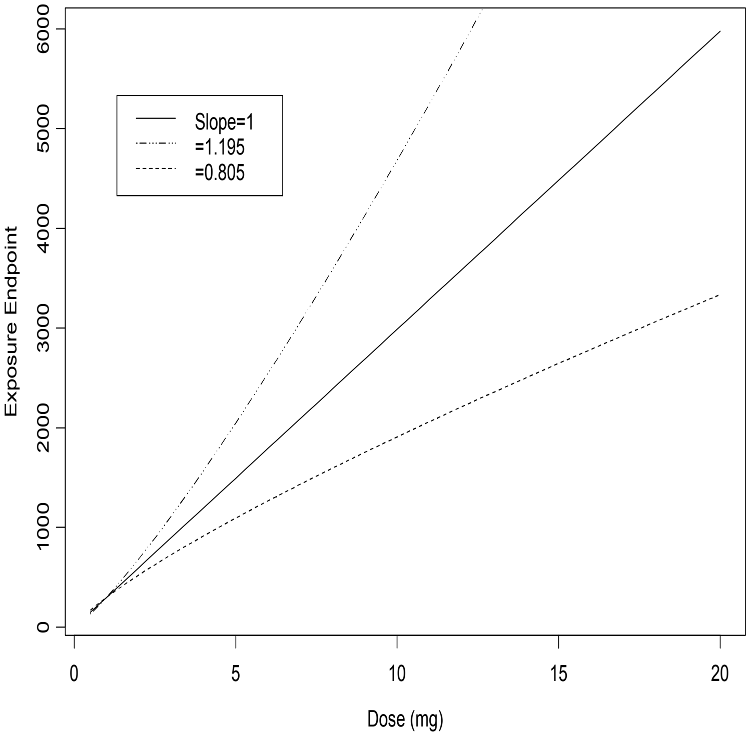

Figure 7.5 *Dose to Exposure (AUC or Cmax) Relationship for β from 0.8 to 1.2*

versus

$$H_{12} : \beta < 1 + t.$$

similar to the TOST used in bioequivalence testing.

However, there is currently no set regulatory standard for the equivalence region. Smith et al. [406] recommends that t be defined as

$$t = \ln \theta / \ln r$$

where θ is the minimal change in exposure beyond which one may want to adjust to maintain safe exposure levels, and r is the ratio of the maximum tolerated or effective dose to be used in the study to the minimum effective dose.

In the following example $\theta = 1.5$, as it was felt for this drug that a 50% increase in exposure might necessitate a decrease in dose. The therapeutic dose range was 1 - 8 mg, and $r = 8$ accordingly. Therefore

$t = 0.195$, and the hypotheses to be tested were:

$$H_{01} : \beta \leq 0.805$$

versus

$$H_{11} : \beta > 0.805$$

and

$$H_{02} : \beta \geq 1.195$$

versus

$$H_{12} : \beta < 1.195.$$

When the parameter β lies between 0.805 and 1.195 (with sufficient confidence), this procedure judges the data adequate to support a claim of dose-proportionality.

As with bioequivalence testing, a mixed model is used to assess the magnitude of β and to derive 90% confidence intervals. If the 90% confidence interval for β lies within $1 - t$ to $1 + t$, for both AUC and Cmax, dose-proportionality is demonstrated.

In Example 7.6.1, a randomized cross-over study in 28 normal healthy volunteers was performed to assess dose-proportionality and the effect of food. SAS code to analyse such data are similar to that applied in Chapter 3 and may be found below.

Table 7.18: Example 7.5.1: AUC and Cmax Data from a Randomized Dose-Proportionality Cross-over Study

Subject	Seq	AUC	AUC	AUC	Cmax	Cmax	Cmax
		A	B	C	A	B	C
1	DCAB	352	746	3408	66.6	208.4	687.2
4	BACD	440	842	2560	88.9	162.6	504.0
5	CBDA	249	552	2856	66.7	124.0	601.6
6	DCAB	318	628	2560	68.9	114.4	495.2
7	ADBC	528	814	3888	98.5	177.8	826.4
8	BACD	512	1122	4680	82.8	204.8	684.8
9	DCAB	329	750	2720	67.0	180.0	510.4
10	ADBC	374	688	2432	65.7	142.8	448.0
11	CBDA	282	994	4680	76.4	191.0	586.4
13	BACD	324	674	2584	82.1	168.8	610.4
14	CBDA	284	636	3176	61.5	108.0	532.0
15	ADBC	372	666	3200	82.8	169.4	792.0
16	DCAB	304	578	2272	67.1	123.8	440.8
17	CBDA	171	400	1696	48.0	90.2	463.2
18	DCAB	489	1054	3752	91.5	190.6	735.2
20	ADBC	267	526	1896	59.9	141.0	540.8

A=1mg; B=2mg; C=8mg; D=8mg with a meal

Table 7.18: Example 7.5.1: AUC and Cmax Data from a Randomized Dose-Proportionality Cross-over Study

Subject	Seq	AUC A	AUC B	AUC C	Cmax A	Cmax B	Cmax C
21	ADBC	292	620	2392	65.6	107.6	332.8
22	BACD	299	580	2488	79.2	126.8	649.6
23	DCAB	392	918	3152	64.1	291.0	615.2
24	CBDA	363	646	3448	87.3	177.2	715.2
25	ADBC	728	896	3232	75.2	130.6	571.2
27	CBDA	348	806	3360	75.5	131.0	560.8
28	DCAB	287	568	2440	69.7	146.2	578.4
29	BACD	283	620	2320	79.4	151.2	502.4
30	CBDA	246	590	2472	78.2	87.6	637.6
31	ADBC	429	786	3264	114.1	186.0	785.6
32	BACD	308	704	2616	81.0	155.2	671.2
33	BACD	462	1132	3656	85.7	174.4	656.8
A=1mg; B=2mg; C=8mg; D=8mg with a meal							

Dose-Proportionality Assessment Example 7.5.1 - SAS `proc mixed` *code:*

```
proc mixed method=reml data=pk_dp;
    class subject sequence period;
    model lnauc=sequence period lndose/
    s ddfm=kenwardroger cl alpha=.1;
    random intercept/subject=subject(sequence);
    run;

proc mixed method=reml data=pk_dp;
    class subject sequence period;
    model lncmax=sequence period lndose/
    s ddfm=kenwardroger cl alpha=.1;
    random intercept/subject=subject(sequence);
    run;
```

The estimates for β were 1.0218 and 0.9879 for logAUC and Cmax, respectively, with 90% confidence intervals contained well within 0.805 to 1.195. Therefore, dose-proportionality was demonstrated. Interested readers may find data for Tmax from this study and AUC and Cmax data for the assessment of food effects (Regimen D compared to Regimen C) on the website accompanying this book. Tmax was not significantly changed by altering the dose of drug, and food did not affect the AUC and Cmax of this drug.

Determination of sample size for such a cross-over study is similar to the procedure used in bioequivalence testing. It is known that the estimate $\hat{\beta}$ is normally distributed with mean β and variance $\frac{\sigma_W^2}{n\sum_m (ld_m - \bar{ld})^2}$, where n is the sample size and $\sum_m (ld_m - \bar{ld})^2$ is the corrected (for the mean logDose, $\bar{ld}$) sum of squared logDoses corresponding to the design matrix of the study. Construction of the TOST procedure in this setting follows from these well-known findings. For the purposes of this calculation, we assume that a design has been selected such that period effects are not related to dose (i.e., that the study is fully randomized) for simplicity.

Sample Size Code for TOST in Dose-Proportionality Studies:

```
data a;
    * total number of subjects
    (needs to be a multiple of number
    of sequences, seq);
    * p is the number of periods;
n=12; seq=3; p=3;
    * significance level;
a=0.05;
    * true dose proportionality;
beta=1;
    * sigmaW = within-subject standard deviation;
sigmaW=0.25;
    * css is the corrected sum of squares of doses;
    * s assumes period effects are orthogonal to dose;
css=CSS(log(1),log(2),log(8));
s=sigmaW/sqrt(n*css);
* error degrees of freedom for cross-over
    with n subjects in total
    assigned equally to seq sequences;
n2=(n*p)-(n+p-1)-1;
    * t = acceptance limit;
    theta=1.25;
    r=8/1;
    t=log(theta)/log(r);
    run;

data b; set a;
* calculate power;
    t1=tinv(1-a,n2); t2=-t1;
    nc1=(sqrt(n))*((beta-(1-t))/s);
    nc2=(sqrt(n))*((beta-(1+t))/s);
    df=n2;
    prob1=probt(t1,df,nc1);
    prob2=probt(t2,df,nc2);
    answer=prob2-prob1;
    power=answer*100; run;
proc print data=b; run;
```

These are typically very powerful designs for the assessment of dose-proportionality, and interested readers will find that power for the above design approaches 100%. Although as few as six normal healthy volunteers will serve to provide a very robust dose-proportionality assessment in most settings, it is recommended that cross-over studies supporting a regulatory file include at least 10 to 12 subjects to ensure application of the central-limit theorem is appropriate.

7.6 Technical Appendix

This technical appendix provides an example of interactive Bayesian modelling of pharmacokinetic data in a first-time-in-humans trial. In Table 7.19, mean AUC and Cmax estimates from a preclinical species are presented.

Table 7.19 *Exposure Estimates from a Preclinical Species*

Dose	Estimated AUC	Estimated Cmax
5 mg/kg	2790	880
100 mg/kg	29,600	7600

Techniques to use these values to predict human AUC and Cmax are discussed in Chapter 30 [12] and will not be discussed further here. For the purposes of illustration, here it is assumed that only human weight needs be taken into account in predicting human exposure levels, and these estimates (assuming a 50 kg human) are provided in Table 7.20.

Table 7.20 *Exposure Estimates for a 50 kg Human from a Preclinical Species*

Dose	Estimated AUC	Estimated Cmax
5 mg	139,500	44,000
100 mg	1,480,000	380,000

We wish to use these data to derive estimates for α and β as discussed in Section 7.2; however, at this stage we have these two unknown parameters and only two data points. For pharmacokinetic data, it is possible to make the assumption that when dose is very small (0.0001) the resulting AUC or Cmax will be very small (0.0001). This yields three data points for two unknown parameters, and a simple regression may be performed to provide prior distributions for α and β. SAS code to perform this analysis may be found on the website accompanying this book. Other means (e.g., expert elicitation) may also be used to derive such estimates for α and β, and we refer interested readers to an excellent review in [154].

In this case, it is estimated that $\hat{\alpha} \sim N(3.43, 0.52)$ and $\hat{\beta} \sim N(1.36, 0.0081)$ where N denotes the normal distribution with (mean, variance) from a regression of logAUC on logDose. We utilize the mean estimates for these parameters in the code below, but assume that the variance associated with them is very wide (reflecting the uncertainty inherent to allometric scaling calculations).

Example 7.2.2 AUC data were analyzed in WINBUGS using the following computer code (based on the code from the RATS WINBUGS example):

Interactive Bayesian First-time-in-humans WINBUGS Analysis Code for Example 7.2.2

```
model
    {for( i in 1 : N )
    {
    for( j in 1 : T )
    {
    Y[i , j] ~ dnorm(mu[i , j],tau.c)
    mu[i , j] <- alpha[i] + beta[i] * x[j]
    }
    alpha[i] ~ dnorm(alpha.c,alpha.tau)
    beta[i] ~ dnorm(beta.c,beta.tau)
    }
    tau.c ~ dgamma(0.001,0.001)
    sigma <- 1 / sqrt(tau.c)
    alpha.c ~ dnorm(3.43,1.0E-6)
    alpha.tau ~ dgamma(0.001,0.001)
    beta.c ~ dnorm(1.36,1.0E-6)
    beta.tau ~ dgamma(0.001,0.001)
    lnmtd <- (7.78-alpha.c)/(beta.c)
    mtd <- exp(lnmtd)
    }
```

This model may be used interactively as data are collected to estimate individual responses (monitoring mu[i , j]) and the MTD. Data and initial values to run this program in WINBUGS may be found on the website accompanying this book.

As with the original analysis, attention is focused on the MTD relative to the NOAEL (2400 for illustration purposes). The MTD in this analysis is estimated as 13.8 mg (using the median posterior density of 100,000 iterations after a burn-in of 1000 iterations). A Bayesian 90%

confidence interval for the MTD is 8.4 to 25.7 mg. Similar analyses may be performed for Cmax, and this is left as an exercise for the reader.

CHAPTER 8

QTc

Introduction

No one can be expected to pay 100% attention to 100% of the issues and data encountered in clinical pharmacology 100% of the time, so one should be forgiven for not recognizing immediately that QTc is a critical issue in drug development.

My boss stepped in one day to alert me to the fact that I now had a new project. We were developing a drug for anti-arrythmia. There were a number of ongoing clinical pharmacology trials that were delivering data, and results would be needed 'Stat' to enable the company to make an investment decision.

I was used to this by this time. No one ever came by and said we had plenty of time to get a job done, with no rush, and that senior management was happy to wait as long as we needed to get the job done at our convenience. I was hopeful at that time that maybe one day I would get a project like that, but now I have given up hope that such an event will ever happen.

In any event, arrythmia denotes an irregular heartbeat. Some are benign, but some are fatal, and the drug we were developing was intended to prevent its occurrence. To do so, my boss informed me that the drug would impact the ECG. I nodded sagely, and after she left I looked it up in my trusty medical dictionary. ECG denotes an electrocardiogram - a tracing of the electrical activity of the heart over time (we will see a typical one later in this chapter). What I was expecting when the data came in, therefore, was a lot of ECG tracings from which I would measure amplitude, trough to trough time intervals, and other summary measures to statistically describe the activity following dosing with our drug relative to placebo. These would obviously be related to the aortas and ventricles I remembered from 8th grade anatomy, so this should not have been too bad.

What I received, however, was a data set of alphabet soup with numbers. There were measurements taken for PR, QRS, QT, RR, QTc, QTcB, QTcF, QTcI (to name a few) in addition to text fields describing T-wave morphology. All of these were measured in triplicate following dosing with placebo and our drug in a pretty large number of patients at many times over the course of a day. There was not an ECG to be seen, *nor any ventricles. It was a completely unidentifiable mass of unbelievable gobbledygook seemingly produced by a team of junior medics with slide rules, protractors, and way too much time on their hands. I found out later it was done by senior medics and had been done this way since the 1920s.*

My guess (which turned out later to be correct) was that these endpoints (PR, etc.) were measuring time relative to the voltage of the heart. But in this instance my medical dictionary let me down. QTc was not in there.

This left me with three options to try to figure out what was going on:

1. *Ask my statistical colleagues (they did not know, or said they did not).*
2. *Go downstairs and talk to Denny about what this stuff was (since the report was needed yesterday), but the problem with talking to Denny was that he would want to know about the data for a couple other projects I was working on, and I did not want to field twenty questions when all I needed was one answer.*
3. *Go to lunch.*

After lunch, I talked to Denny and got a crash course on the heart and electrocardiology. QTc turned out to be very important, not only for this drug, but also as a general issue in drug development. We will devote this chapter to QTc as it is now an important issue assessed in clinical pharmacology assessments of drug safety for all drug products.

8.1 Background

An electrocardiogram (ECG) measures the electrical activity of the heart over time. Usually, eight 'leads' or electrical monitors are placed on a patient's upper torso and back along certain predetermined vectors out from the heart. These leads then monitor the electrical output of the heart to construct a graph of the polarization and depolarization of different parts of the heart during a beat. See Figure 8.1. This pattern is repeated over and over again while the heart beats.

The different parts of the ECG are denoted by letters and referred to as 'waves' and 'complexes'. For instance, the first 'bump' is referred to as the P-wave. The nadir of the first dip begins the QRS-complex, and the wave immediately following this complex is the T-wave. In some ECGs, there is a following wave known as the U-wave, but this is unusual in normal healthy volunteers.

On the ECG tracing, the QT interval is defined as the amount of time between the initiation of the QRS complex and the conclusion of the T-wave. QT interval duration is measured in milliseconds (msec), by computer algorithm, and measures of how long it takes the heart to

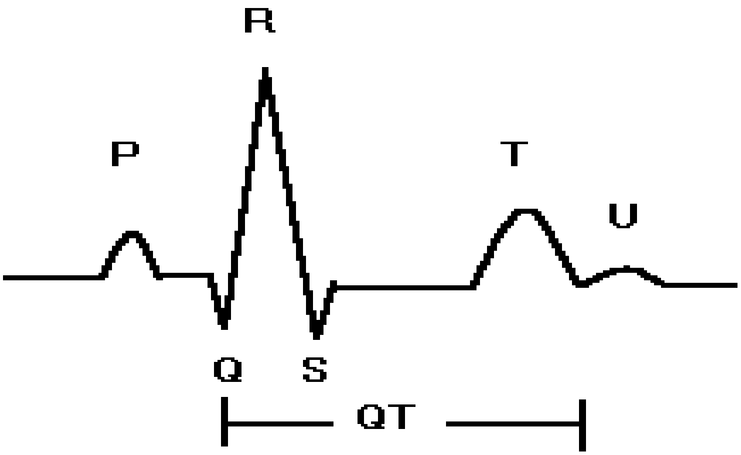

Figure 8.1 *A Typical 12-Lead ECG Interval*

repolarise and prepare for its next beat. The longer it takes to repolarise, the more time between beats, and the less oxygen gets to cells.

QT duration is dependent upon gender, age, health status, menstrual cycle, and a great number of other factors. QT changes naturally over the course of the day, and QT duration can be prolonged by food and is changed by exercise. Some drugs prolong the QT interval (i.e., delay the heart's ability to repolarise). If QT is prolonged sufficiently in humans, potentially fatal cardiac arrhythmias can result. Torsades de Pointes, most often referred to in connection with QT as these are known to be related, is a malignant ventricular arrhythmia known to occur infrequently in individuals who are genetically predisposed to this condition and sometimes in response to drug therapy [349].

Prolongation of the QTc interval has been observed to be related to increased risk of Torsades de Pointes in an exponential fashion [310]. The QT interval is highly correlated with how fast the heart is beating overall (measured by determining RR, the length of time between one R on the ECG and the next R). Therefore in measuring QT, the interval is usually corrected to derive a QTc (QT interval corrected for heart rate). Common corrections were developed by Fridericia [149] and Bazett [23], and many authors have published on better ways to correct for heart rate in recent years, e.g., [91]. Bazett's correction has been observed to overcorrect QTc at some heart rates [353], [422], [336] and is not generally used for the purposes of safety assessment described in this chapter. We will not dwell further here on the application of correction factors in this setting, and will utilize Fridericia's correction (QT is corrected by division of the cube-root of RR such that $QTcF = \frac{QT}{RR^{1/3}}$) in subsequent discussion as it appears unrelated to heart rate according to recent reports [353], [422].

QTc prolongation is a necessary but not sufficient condition for occurrence of and has a qualitative relationship to clinical arrhythmias [229]. One must, by definition, have a prolonged QTc just prior to the occurrence of Torsades de Pointes, but a prolonged QTc can occur without the occurrence of Torsades de Pointes. In general, a prolonged QTc in a patient with several other risk factors [5] may result in Torsades de Pointes. Prolongation from baseline (usually taken first thing in the morning) in an individual greater than 60 msec or an absolute value of QTc beyond 500 msec is deemed a clinical safety signal [229].

Drugs known to prolong the QTc interval have been responsible for killing people. This potential was observed for Terfenadine [213], [346]-[347], Cisapride [462], and other examples [426]. Terfenadine and Cisapride were approved and marketed compounds when the deaths due to drug occurred. The potential for this effect was identified only after the drug was marketed to a large number of patients, and these and several other drugs were withdrawn from the market to protect patient safety [426]. This highlighted the need for thorough assessment of the potential for QTc prolongation prior to approval.

New drugs, and potentially existing drugs seeking new indications, must study and rule out the potential for prolongation of QTc [229]. This thorough study will rule out the presence of a QT/QTc prolongation, or inform how much monitoring for QTc potential will be necessary to establish safety to market in confirmatory trials. Mean prolongation of QTc in excess of 5 to 8 msec will merit greater scrutiny in confirmatory trials. Prolongation greater than 20 msec will likely result in refusal to market unless the benefit of the drug product far outweighs the risk of QTc prolongation and clinical arrythmia (e.g., for an oncology agent).

Even if such a product were approved, it would likely have stringent warnings and requirements limiting its use to patients where benefit clearly outweighs risk. However, such labelling has been observed to be ineffective in the past at protecting patients in the marketplace [426].

Now that the reasons behind assessment of QTc prolongation have been developed, we turn to discussion of how to model data from a thorough QTc study. This will be followed by a section on design of thorough QTc studies, and last we will consider how to interpret the results of such trials.

8.2 Modelling of QTc Data

To illustrate an approach to the modelling of QTc data, we will consider some data from previous trials. Three cross-over data sets were selected for use as examples in this chapter. These example data sets were selected to provide a range of example effect sizes in QTc prolongation relative to changes in ECG sampling procedures and sample sizes. Normal healthy volunteers are generally [229] the population dosed in such studies as it is felt that QTc prolongation observed in that population does not pose a great risk, and findings are readily applicable to patient populations.

Fridericia's correction to the QT interval was used in all analyses, and all studies were fully randomized cross-over designs in normal healthy volunteers. In all cases, the objective of the trial was to detect changes in QTc induced by study drug over and above those introduced by a control agent, and ECGs were manually over-read by a qualified, blinded cardiologist.

In our first example data set (Example 8.1, SAS data set `exam8_1`), three single-dose regimens (C, D, E) were studied relative to placebo control (Regimen F). Regimen E was a known mild prolonger of the QTc interval (included to serve as a positive control), and regimens C and D were a therapeutic and supra-therapeutic dose of a moderate QTc prolonging agent. Forty-one subjects were included in the example data set, and QTc was measured in triplicate at baseline (time 0) and over the course of the day at set times following dosing. Triplicate (three ECGs) measurements were averaged at each time of ECG sampling (i.e., 0, 0.5,1, 1.5, 2.4, 4, etc.) for inclusion in analysis, and samples out to four hours post dose were included in the example data set for ease of presentation and discussion.

In the second example data set (Example 8.2, SAS data set `exam8_2`), two seven-day, repeat-dose regimens were studied (Regimens A and C) relative to a seven-day, repeat-dose regimen of placebo (Regimen F). Regimen A was a known severe prolonger of the QTc interval, and it was of interest to study whether this drug in combination with another (a metabolic inhibitor of the drug of interest, denoted Regimen C) would result in even greater prolongation. Twenty-three subjects were included in this example.

In the last data set (Example 8.3, SAS data set `exam8_3`), a seven-day, repeat-dose regimen of the combination of Terfenadine with a potential metabolic inhibitor (Regimen B) was studied relative to Terfenadine alone (F). Eleven subjects were included in this example.

Consider some of the first subject's QTc data as listed in Table 8.1 of Example 8.1 below.

Unlike bioequivalence, where only one AUC or Cmax observation was

Table 8.1 *First Subject's Data in Example 8.1*

Subject	Regimen	Time(h)	QTc(msec)
1	C	0.0	358
1	C	0.5	356
1	C	1.0	361
1	C	1.5	362
1	C	2.5	354
1	C	4.0	355
1	D	0.0	373
1	D	0.5	381
1	D	1.0	389
		......	

of interest in each period, as with the safety data discussed in Chapter 7, in this setting the pattern of QTc response within and across periods is of interest. Such repeated-measures, time-series data are inherently more complex to model. However, many methods are available to do so.

The analysis of such repeated measures data arising in cross-over studies with baseline control is described in Jones and Kenward, Ch5 [237]. This analysis accounts for each subject as their own control, the correlation between measurements within-period, and accounts for baseline, period, and regimen effects. The SAS code one may use to do so follows.

```
proc mixed data=for_an method=reml
    CL scoring=50 maxiter=200;
    class subject period rel_time regimen;
    model qtcf=qtcfb period regimen rel_time
    period*rel_time regimen*rel_time
    /DDFM=KENWARDROGER S outp=out;
    random subject;
    repeated rel_time/type=AR(1)
    subject=subject*period;
    lsmeans rel_time*regimen/corr cov;
    ods output LSMeans=means; run;
```

As in the earlier examples, `proc mixed` is called in SAS, and told to use REML modelling, and to do a maximum of 200 iterations (the `maxiter` statement). The `class` statement describes the descriptor variables of the data set appropriate to the cross-over design, and `rel_time` denotes the time of ECG sampling from which QTc was measured relative to dosing at time 0 hours.

The model statement indicates that baseline QTc (`qtcfb`), period, regimen, and time of measurement (along with appropriate interactions of these terms) should be assessed for impact on fitted model. The `random` statement indicates that each subject should be treated as their own control, and the repeated statement specifies that the values of `rel_time` should be regarded as correlated within each subject's period with the correlation decreasing with increasing duration between times of ECG sampling.

To describe the pattern of overall response to treatment, the model-adjusted means are output (along with their correlation and variance-covariance matrix) in the `lsmeans` and `ODS` statements.

The `mixed` procedure accounts for effects as described above, and provides adjusted mean estimates for use in describing the average effect of treatment. These are plotted for Examples 8.1-3 in Figures 8.2 to 8.4 below.

The adjusted mean estimates derived from the mixed model are known as 'BLUP' in that they are denoted as *Best Linear Unbiased Predictors*. They are asymptotically unbiased estimators for the behavior of mean QTc in the population being studied, and as with bioequivalence, will serve to compare the properties of the different treatments.

In Figure 8.2, mild (Regimen E) and moderate degrees of prolongation (Regimen C) relative to Regimen F (placebo) are observed with slightly greater prolongation being observed at the supra-therapeutic dose of the drug being studied (Regimen D). In this context, we denote 'mild' as referring to a QTc prolongation that does not begin until some time after a dose of drug is administered and which rapidly dissipates over time. 'Moderate' QTc prolongation in contrast denotes a QTc prolongation that begins rapidly after a dose is administered and is maintained over a substantial part of the dosing interval. Both mild and moderate prolongation refer to effect sizes greater than zero but less than the ICH E14 [229] level of probable concern for causing Torsades de Pointes of 20 msec [426].

In Example 8.2, plotted in Figure 8.3, severe QTc prolongation is observed in response to treatment in Regimens A and C. Some evidence of a synergistic effect (Regimen C) appears present, and we will evaluate whether this effect is statistically differentiable later in the chapter. 'Severe' QTc prolongation in this example denotes mean QTc prolongation that is equal to or in excess of the ICH E14's [229] stated level of clinical concern as probably leading to Torsades de Pointes (20 msec) at any time following a dose of drug as compared to placebo.

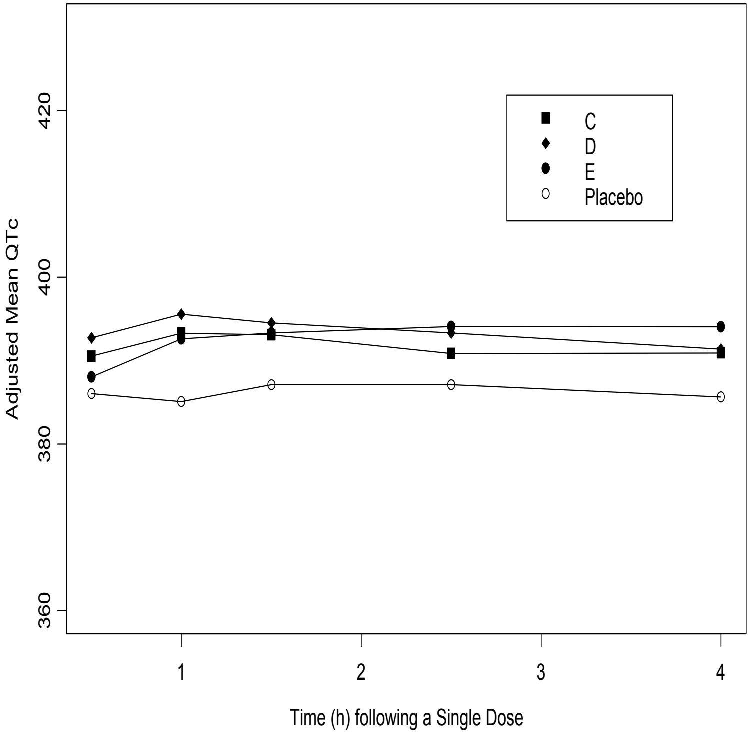

Figure 8.2 *Mild and Moderate QTc Prolongation* ($n = 41$) *in Example 8.1*

In Example 8.3, however, plotted in Figure 8.4, no clear evidence of QTc prolongation is observed in response to treatment in Regimen B.

As we begin considering statistical methods to compare these responses, we should consider one other important issue in the modelling of repeated measures data. That is, that the mean responses within a regimen and across regimens are correlated given the nature of the cross-over study design and repeated measures ECG data. For example, consider the model-adjusted means for Regimen F (placebo, accounting for all other model parameters) in Example 8.1 (see Table 8.2). In Table 8.2, the entries of ρ_i denote the estimated correlation between the adjusted mean at time i and the other means at times $j \neq i$.

For example, the adjusted mean QTc at time 0.5 is perfectly correlated with itself ($\rho_{0.5} = 1$, as one would expect) but is correlated to a lesser extent with the adjusted mean at time 1 ($\rho_1 = 0.53$), and is correlated to an even lesser extent with the adjusted mean at time 1.5 ($\rho_{1.5} = 0.34$).

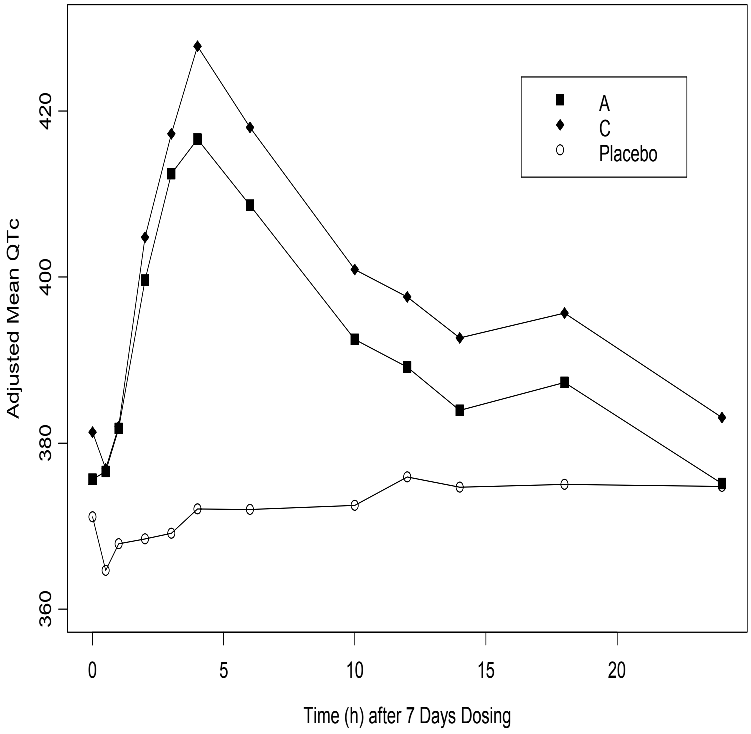

Figure 8.3 *Severe QTc Prolongation* (*n* = 23) *in Example 8.2*

This indicates that if we know the response at time 0.5, we have some indication (as measured by ρ) of what the response at 1 hour will be (and vice versa). As the time interval increases between ECG collections, the correlation decreases.

If these means were independent, the estimated ρ would be close to null. As these adjusted means are derived from a cross-over trial, the adjusted means between regimens are also correlated. Therefore, when we begin comparing treatments we can account for the fact that adjusted means between treatments are correlated and that adjusted means are also correlated across time.

This has implications for statistical testing as when we test for differences between regimens at one particular time of ECG sampling, this gives us an indication of what we will find at other times of sampling. Ignoring this relationship in statistical calculations leads to narrower confidence intervals. This makes the control of Type 1 and 2 error in

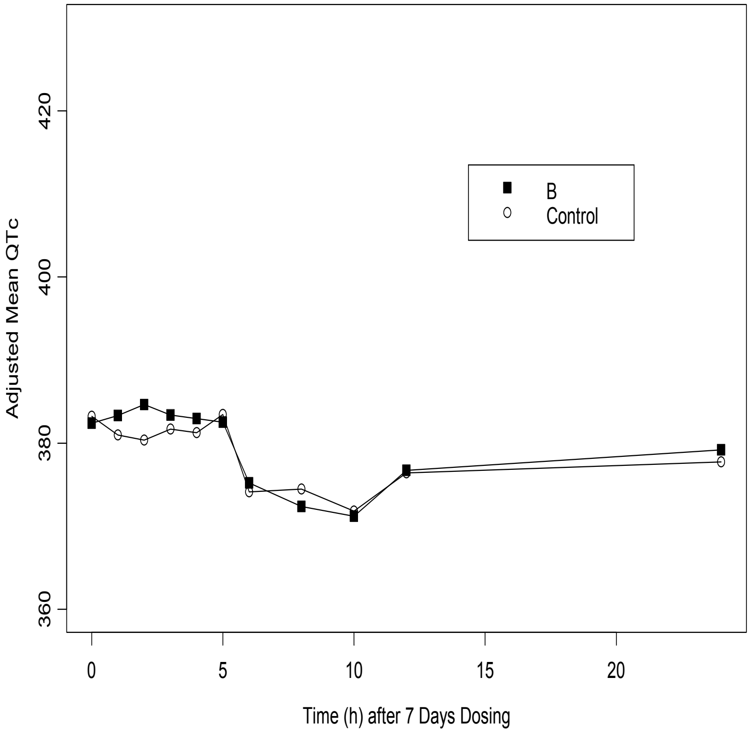

Figure 8.4 *No QTc Prolongation (n* $= 11$*) in Example 8.3*

statistical testing somewhat more complicated, and we will discuss some of the implications of this later in this chapter.

SAS `proc mixed` computes comparisons between treatments quite simply. In the `mixed` code above, the lsmeans statement is replaced with

```
lsmeans regimen*rel_time/DIFF CL ALPHA=0.1;
```

to construct all possible differences between each adjusted mean QTc (with 90% confidence intervals from the options `ALPHA=0.1` and `CL`. These may be output to a data set `diffs` by the addition of `Diffs=diffs` to the `ods` statement.

As we are most interested in the comparison between treatments at each time of ECG collection, the non-relevant comparisons may be eliminated by means of a data step where the comparisons where times differ are eliminated by:

Table 8.2 *Adjusted Mean QTc (msec) and Estimated Correlation Matrix for Regimen F of Example 8.1*

Regimen	Time(h)	AM.QTc	$\rho_{0.5}$	ρ_1	$\rho_{1.5}$	$\rho_{2.5}$	ρ_4
F	0.5	386.05	1.0000	0.5342	0.3383	0.2559	0.2213
F	1	385.09	0.5342	1.0000	0.5342	0.3383	0.2559
F	1.5	387.09	0.3383	0.5342	1.0000	0.5342	0.3383
F	2.5	387.09	0.2559	0.3383	0.5342	1.0000	0.5342
F	4	385.63	0.2213	0.2559	0.3383	0.5342	1.0000

F = Placebo

```
if REL_TIME ne _REL_TIME then delete;
```

These comparisons account for the correlation between regimens at each individual time (i.e., that Regimen C is correlated with Regimen F at time 0.5 for example); however, they do not account for the correlation between means across the time interval of ECG sampling (that the means at time 0.5 h are correlated with the means at time 1 h, etc.). It is up to the user, however, to determine which means of controlling the Type 1 and 2 error rates should be employed, and SAS does not automatically do so. To begin discussion on this topic, we first consider the results (not adjusted for correlation across time) as presented in Tables 8.3 to 8.5.

In Example 8.1, it is observed that moderate and statistically significant (note lower 90% confidence bounds exceed zero) QTc prolongation is observed in Regimens C and D within a half-hour of dosing and remains prolonged out to four hours post dosing. Significant prolongation for regimen E is not observed until after a half-hour following dosing and returns to parity with Regimen F immediately after four hours post dose (data not shown).

Following seven days of dosing in Example 8.2, significant and severe prolongation begins within a half-hour of a dose of Regimen A relative to control. This continues throughout the day and returns to parity with control 24 hours later. In contrast, when the drug of interest is given with a metabolic inhibitor (Regimen C), significant prolongation is observed prior to the last dose on day seven being given (comparison C-F at time 0h), and QTc continues to be prolonged even 24 hours later.

In the last Example 8.3, however, no statistically significant changes in Regimen B were detected relative to control after seven days of dosing.

Now that models and simple statistical procedures to compare data between regimens have been developed, we now discuss statistical procedures to control the Type 1 and 2 error rates in testing for QTc prolongation.

Table 8.3 *Mean Changes (90% CI) Between Regimens Following a Single Dose in Example 8.1 (n* = 41*)*

Comparison	Time	Difference	90%CI
C-F	0.5	4.4923	(2.1997,6.7848)
	1	8.1830	(5.8904,10.4755)
	1.5	6.0120	(3.7195,8.3045)
	2.5	3.7444	(1.4518,6.0369)
	4	5.2944	(3.0018,7.5869)
D-F	0.5	6.6868	(4.4035,8.9701)
	1	10.4591	(8.1758,12.7425)
	1.5	7.4421	(5.1588,9.7255)
	2.5	6.2212	(3.9379,8.5046)
	4	5.7591	(3.4757,8.0424)
E-F	0.5	2.0069	(-0.2778,4.2915)
	1	7.5171	(5.2324,9.8017)
	1.5	6.2216	(3.9369,8.5062)
	2.5	6.9994	(4.7147,9.2840)
	4	8.4446	(6.1599,10.7292)

C = Therapeutic Dose
D = Supra-therapeutic Dose
E = Dose of Positive Control
F = Placebo

8.3 Interpreting the QTc Modelling Findings

As with bioequivalence testing, in the context of QTc testing we are interested in confirming that a difference in treatments is **not** present. It is presumed that the drug of interest does prolong QTc until it is demonstrated not to be the case. The hypotheses of interest are therefore similar to bioequivalence testing, and the burden of proof remains on the sponsor of the study to demonstrate that QTc prolongation does not occur.

Here we are generally interested in confirming that QTc is not prolonged following dosing over an appropriate period of time when the drug could cause such an effect. The model estimates described in the last section will be used to test for an effect over the appropriate ECG sampling interval. In this setting the null hypothesis for comparison of the test drug (T, at either the therapeutic or supra-therapeutic dose)

Table 8.4 *Mean Changes (90% CI) Between Regimens Following Seven Days of Dosing in Example 8.2 ($n = 23$)*

Comparison	Time	Difference	90%CI
A-F	0	4.5452	(-1.2812,10.3716)
	0.5	11.9137	(6.0873,17.7401)
	1	13.8814	(8.0550,19.7078)
	2	31.1529	(25.3265,36.9793)
	3	43.3057	(37.4793,49.1321)
	4	44.5157	(38.6893,50.3421)
	6	36.6313	(30.8049,42.4577)
	10	19.9793	(14.1529,25.8057)
	12	13.2215	(7.3951,19.0479)
	14	9.2482	(3.4218,15.0746)
	18	12.2382	(6.4118,18.0646)
	24	0.3131	(-5.5133,6.1395)
C-F	0	10.2451	(4.4200,16.0703)
	0.5	12.2385	(6.4133,18.0637)
	1	14.1211	(8.2959,19.9463)
	2	36.3119	(30.4867,42.137)
	3	48.1345	(42.3093,53.9597)
	4	55.7178	(49.8926,61.5430)
	6	46.0229	(40.1977,51.8480)
	10	28.4315	(22.6063,34.2566)
	12	21.6648	(15.8396,27.4900)
	14	17.9892	(12.1640,23.8143)
	18	20.6361	(14.8109,26.4613)
	24	8.3007	(2.4756,14.1259)

A = Severe Prolonger
C = Severe Pro. with Metabolic Inhibitor
F = Placebo

relative to placebo (P) is:

$$H_0 : \mu_{Ti} - \mu_{Pi} \geq \Delta \tag{8.1}$$

for at least one i where i denotes ECG samples collected over the relevant times of sampling and Δ is a predetermined, reasonable no-effect goalpost (defined in [229]). This hypothesis is to be tested versus the

Table 8.5 *Mean Changes (90% CI) Between Regimens Following Seven Days of Dosing in Example 8.3 ($n = 11$)*

Comparison	Time	Difference	90%CI
B-F	0	-0.8319	(-7.5455,5.8817)
	1	2.3494	(-4.3642,9.0630)
	2	4.2809	(-2.4327,10.9945)
	3	1.7183	(-4.9953,8.4319)
	4	1.7070	(-5.0066,8.4206)
	5	-0.9137	(-7.6273,5.7999)
	6	1.0656	(-5.6480,7.7792)
	8	-2.0849	(-8.7985,4.6287)
	10	-0.5964	(-7.3100,6.1172)
	12	0.3033	(-6.4103,7.0170)
	24	1.4409	(-5.2727,8.1545)

B = Terfenadine with Metabolic Inhibitor
F = Terfenadine alone

alternative hypothesis:

$$H_1 : \mu_{Ti} - \mu_{Pi} < \Delta \tag{8.2}$$

for all i in the sampling interval. This type of testing procedure is not new. It is sometimes referred to as an intersection-union test and is described in detail in Wellek (Chapter 7) [447].

Performing this test is simple and straightforward. One simply derives the comparisons of interest between treatments using SAS as described in the last section (Tables 8.3 to 8.5) and evaluates the magnitude of the upper bound of the 90% confidence intervals over the relevant interval of sampling relative to the chosen Δ. The level of Δ in [229] was a subject of much debate when [229] was being developed; however, if we choose either 10 msec for discussion purposes, we see that the null hypothesis is not rejected for the mild and moderate QTc prolongers of Example 8.1, is clearly not rejected for the severe prolonging treatments of Example 8.2, and is not rejected for Regimen B of Example 8.3.

This last finding is surprising on the surface, but is not terribly unexpected given the nature [447] of the testing procedure being used. As we know from bioequivalence testing, Type 1 and 2 errors can occur in testing of such situations.

		The Truth	
		Trt is NOT Safe	Trt IS Safe
Statistics from study show that	Eqn (8.1) NOT Rejected (Trt NOT Safe)	Right answer!	Wrong answer (Type 2 error)
	Eqn (8.1) IS Rejected (Trt IS Safe)	Wrong answer (Type 1 error)	Right answer!

The findings of Example 8.3 are an example of a Type 2 error driven by the limited sample size ($n = 11$). To prove that a treatment is safe under this approach, it must be shown that it is safe over the entire sampling interval. The intersection-union test is known to protect against Type 1 errors at a very conservative level (i.e., less than or equal to the desired level of 5%). This makes the occurrence of a Type 1 error infrequent, a desirable property of such a testing procedure. The risks associated with admitting a drug to the marketplace that prolongs QTc were discussed in Section 8.1, and it is clear that regulators should be concerned (and conservative) with control of the Type 1 error rate.

The potential for a Type 2 error is best controlled in design. Techniques for doing so are:

1. Increase the sample size (n),
2. Increase the number of ECGs collected at each time point (r),

Both these actions result in smaller confidence intervals about the model estimates of effect, increasing the precision of the study, yielding more confidence in the understanding of the exact properties of the treatment.

In practice a combination of both is done as appropriate to the treatment under study. In Example 8.3, we could most likely prevent the occurrence of a Type 2 error by increasing the sample size to n=30 to 40 subjects or equivalently by increasing the number of ECGs collected at each timepoint from r=1 (as was done in the example) to r=3 to 4 ECGs (working to reject Equation (8.1) with a $\Delta = 10$msec). In working practice, sponsors of such trials get very depressed when a Type 2 error occurs, so they generally do both. More details pertaining to sample size selection and Type 1 and 2 error rates may be found in the Technical Appendix to this chapter.

Note that sample sizes and ECG sampling under this approach also detect even mild and moderate degrees of QTc prolongation as shown in Example 8.1. Severe QTc prolongation may be detected with smaller sample sizes and less stringent ECG sampling as seen in Example 8.2.

ICH E14 [229] goes into a great deal of complex detail on how to demonstrate that a drug does not affect QTc. To rule out that QTc prolongation occurs for a particular treatment, it must be shown that in comparing study drug to placebo at therapeutic and supra-therapeutic doses:

> the largest time-matched mean difference (baseline-subtracted) for the QTc interval is around 5 msec or less, with a one-sided 95% confidence interval that excludes an effect $>$ 10 msec [229].

To perform this procedure one would inspect Tables 8.3 to 8.5 for the maximum difference in adjusted means and then use an appropriate statistical procedure to construct a confidence interval on this quantity, accounting for its correlation to all the other comparisons at other ECG sampling times. Mathematically, this procedure is quite complex. In practice, however, it turns out to be equivalent [447] to the intersection-union test when using the SAS repeated-measures cross-over model described in Section 8.2.

The high degree of correlation in QTc data over time also suggests that the intersection-union testing procedure is not terribly conservative. Estimates of auto-regressive correlation (a measure of how related data are across time for individual subjects) were 0.9 and 0.8 in the Placebo data of Examples 8.1 and 8.2, respectively (a value of 1 would indicate perfect correlation). Correlation was still high (0.7) in the Terfenadine control arm of Example 8.3 - a drug known to prolong QTc. So although slightly conservative in its control of Type 1 errors, intersection-union testing will likely meet regulatory, sponsor, and statistical considerations for testing of safety for the issue of QTc prolongation.

Up to now, we have discussed comparisons of a given treatment with a control. However, in trials performing thorough QTc assessments, it is not unusual for multiple doses and a positive control treatment to be employed [229]. See Example 8.1 where supra-therapeutic and therapeutic doses were employed. As the number of doses increases, so too does the possibility of a Type 1 or 2 error. To control these probabilities, one should follow the principles of proof of safety testing described by Hauschke and Hothorn [199] for this setting [336]. A predefined testing procedure should be used to logically order of statistical tests to mitigate the probability of a Type 1 error. One (step-up) procedure is as follows:

1. Compare the therapeutic dose to placebo. If Equation (8.1) is rejected in favor of Equation 8.2, then the therapeutic dose is acceptable (proof of safety has been demonstrated) and proceed to Step 2; otherwise, stop and conclude that safety has not been demonstrated at the therapeutic dose **and** the supra-therapeutic dose.

2. Compare the supra-therapeutic dose to placebo. If Equation (8.1) is rejected in favor of Equation (8.2) for the supra-therapeutic dose, then the supra-therapeutic dose is acceptable (proof of safety has been demonstrated); otherwise, stop and conclude that safety has not been demonstrated at the supra-therapeutic dose but was at the therapeutic dose.

Another (step-down) procedure is as follows:

1. Compare the supra-therapeutic dose to placebo. If Equation (8.1) is rejected in favor of Equation (8.2) for the supra-therapeutic dose, then the supra-therapeutic dose and the therapeutic doses are acceptable (proof of safety has been demonstrated); otherwise, conclude that safety has not been demonstrated at the supra-therapeutic dose but conduct additional testing at the therapeutic dose by proceeding to Step 2.

2. Compare the therapeutic dose to placebo. If Equation (8.1) is rejected in favor of Equation (8.2), then the therapeutic dose is acceptable (proof of safety has been demonstrated); otherwise, stop and conclude that safety has not been demonstrated at the therapeutic dose **and** the supra-therapeutic dose.

A sequentially rejecting procedure [199] is appropriate for application under the assumption that QTc prolongation increases with dose. This relationship has been observed for most drugs known to prolong the QTc interval [426]. The role of the positive control (if any) is discussed in the next section.

Alternatives to the intersection-union test are available. Such statistical testing procedures control Type 1 error at the precise level of 5% while minimizing the probability of a Type 2 error based upon correlations observed in the data. The Technical Appendix to this chapter describes application of one such technique (Westfall's SimIntervals approach [449] based upon [392] and [80]) using a SAS program available in [450] applied to Example 8.1. As the regulatory acceptance of such an approach is unknown, however, we do not discuss it further here.

8.4 Design of a Thorough QTc Study in the Future

The objectives of thorough QTc studies in the future will be to:

1. Confirm that the new drug does not prolong QTc to a clinically relevant extent (Equation (8.1)), or

2. To measure the extent to which a drug prolongs QTc.

Given the potential risks induced when QTc is prolonged, it is expected that such compounds will likely be screened out of consideration early in drug development, and that in most cases the first objective will be the primary objective of most trials.

In either case, however, QTc will be measured at baseline and over a 24-hour sampling period following dosing with the therapeutic dose of study drug, a supra-therapeutic dose of study drug, placebo, and possibly a positive control (e.g., Moxifloxacin, an antibiotic known to prolong QTc.) As with bioequivalence testing, normal healthy volunteers will generally be used as the study population [229].

When selecting a design to assess Equation (8.1), a cross-over design will likely be the most sensitive and efficient in providing data to assess the null hypothesis. If a drug truly has no effect on QTc, then one would not expect carry-over effects to be an issue in the use of cross-over designs for the trial. If however, there is suspicion that the drug may prolong QTc and the drug has a long-half life, then a parallel group trial may also be used to test Equation (8.1) to avoid the potential for carry-over to confound interpretation of the results.

When using cross-over designs, it should be noted that period effects occur in such thorough QTc assessments, and treatments should be fully randomized throughout the study periods to ensure that period effects are not confounded with treatment effects. Period effects can be induced by small period-to-period differences when using a manual over-reader for the ECGs, and it is expected that computer algorithmic measurement is less prone to such effects.

Computer algorithmic measurement is not perfect, however. It is known to be conservative in its assessment of the end of the T-wave, erring on the side of caution to ensure that individual QTc values of potential concern (QTc $>$ 500 msec) are captured. Computer algorithmic measurement is held [229] to be biased in such assessment individual value assessment; however, as the analyses conducted as described in Sections 8.2 to 8.3 account for baseline QTc and each subject as their own control, it would not be expected that such measurement bias introduced by using computer algorithm would impact statistical inference for Equation (8.1).

Thorough QTc evaluations will generally be conducted in late Phase II or in parallel with the confirmatory trials for regulatory submission. Such trials cannot be conducted unless one has a good idea of the therapeutic and supra-therapeutic doses for the drug of interest, and firm knowledge of these is generally not available until one has demonstrated proof-of-concept and done some work in dose-finding in patients (see Chapter 9).

The choice of Δ has been noted as worthy of further discussion in the draft ICH E14 guidance [229] and has been defined as 8 msec for the purposes of opening discussion. It was originally proposed as 5 msec, then changed to 7.5 in ICH discussions, before taking on the value of 8 msec. Dependence on choice of endpoint was highlighted in [336]. In the final Step 4 guidance [229], 10 msec was defined as the Δ.

Throughout this chapter we have used 90% confidence intervals to describe the QTc data. However, readers will note that Section 8.3's assessment of proof of safety is primarily driven by inspection of only the upper bound of the confidence interval. We have chosen to employ these confidence intervals to recognize that regulation [229] on this topic is imperfect. As discussed previously, the choice of Δ is ill-defined, and QTc is a necessary but not sufficient condition for the development of Torsades de Pointes. As discussed in Chapter 1 (Bernoulli's Principle 8), we should not attribute more weight to such a matter than its due and should view the safety assessments made from the statistics for QTc with some level of caution. ICH E14 [229] is simply a tool being used to protect the public. It is thought that had this been done people would not have died. The reader should recall that as per discussion in Chapter 1, in reality, complete certainty of safety cannot be achieved by such safety testing.

In such a context, the 90% confidence intervals serve a dual purpose. They provide the basis for the QTc safety assessment (using the upper bound), but should mild or moderate prolongation be observed, the lower confidence bound and point estimate serve to place this effect size in context and to evaluate its statistical significance and probability of hazard [199].

A positive control does not add value in a thorough QTc evaluation when testing for a new treatment's safety at therapeutic and supra-therapeutic doses relative to placebo. However, its inclusion does add value if a statistically significant prolongation is observed (i.e., the lower bound of the CI is nonnegative). In drugs known to prolong QTc, the positive control's inclusion in a thorough assessment serves as a method to construct 'no worse than' statistical tests.

Consider the comparison of regimens C and D to E in Example 8.1 and regimen A to C in Example 8.2, as described in Tables 8.6 and 8.7.

In Table 8.6, we see that QTc was prolonged greater than the positive control at the therapeutic and supra-therapeutic doses early in the sampling, appeared similar in the middle of the sampling, and was lower than the positive control at 4 hours post dose. Though we cannot conclude that the new drug poses no risk of QTc prolongation, the statistically significant hazard introduced by dosing with the new drug at supra-therapeutic and therapeutic doses (2 to 5 msec immediately following dosing) does not appear markedly dissimilar to that produced by the positive control later in the day.

Table 8.6 *Mean Changes (90% CI) Between Test Drug and Positive Control Following a Single Dose in Example 8.1 ($n = 41$)*

Comparison	Time	Difference	90%CI
C-E	0.5	2.4854	(0.1957,4.7751)
	1	0.6659	(-1.6237,2.9556)
	1.5	-0.2096	(-2.4992,2.0801)
	2.5	-3.2550	(-5.5447,-0.9653)
	4	-3.1502	(-5.4399,-0.8605)
D-E	0.5	4.6799	(2.3953,6.9646)
	1	2.9421	(0.6574,5.2267)
	1.5	1.2206	(-1.0641,3.5052)
	2.5	-0.7781	(-3.0628,1.5065)
	4	-2.6855	(-4.9701,-0.4009)

C = Therapeutic Dose
D = Supra-therapeutic Dose
E = Dose of Positive Control

In Table 8.7, we see that statistically significant evidence of increased hazard (8 to 11 msec) was observed for the combination of the test drug with a metabolic inhibitor beginning approximately 4 hours after the combination was administered. As dosing of the test drug was already hazardous [229] (see previous section), administration of these two drugs in combination should be viewed with even enhanced caution.

As techniques in clinical pharmacology safety assessment have now been reviewed, we now turn to the assessment of efficacy and mechanism of action for drug products.

8.5 Technical Appendix

8.5.1 Intersection-Union Test (IUT) for No-Effect on QTc

To study the properties of the IUT with regard to assessment of effect on QTc, a data set from a cross-over study was utilised. In this study, normal healthy subjects received placebo repeatedly over 6 weeks, and their QTcF was measured (in triplicate) at baseline (time= 0 hours) and following dosing at 0.5, 1, 1.5, 2, 3, 4, 6, 8, 12, and 24 hours post dose. This is a standard clinical pharmacology sampling scheme, and it is

Table 8.7 *Mean Changes (90% CI) Between Severe QTc Prolonger with and without a Metabolic Inhibitor Following Seven Days of Dosing in Example 8.2 ($n = 23$)*

Comparison	Time	Difference	90%CI
C-A	0	5.7000	(-0.2377,11.6376)
	0.5	0.3248	(-5.6128,6.2625)
	1	0.2397	(-5.6979,6.1774)
	2	5.1589	(-0.7787,11.0966)
	3	4.8288	(-1.1089,10.7664)
	4	11.2021	(5.2645,17.1398)
	6	9.3916	(3.4539,15.3292)
	10	8.4521	(2.5145,14.3898)
	12	8.4433	(2.5057,14.3810)
	14	8.7410	(2.8033,14.6786)
	18	8.3979	(2.4602,14.3355)
	24	7.9877	(2.0500,13.9253)

A = Severe Prolonger
C = Severe Pro. with Metabolic Inhibitor

expected that such a scheme will be utilised in the majority of thorough QTc trials.

The bootstrap was used to generate 2000 bootstrap data sets of size $n = 10$ to 60, and each subject in each bootstrap study was randomly assigned (using `proc plan`) to a sequence of dummy treatments to mimic randomisation which will take place in clinical trials. The model of Jones and Kenward [237] was utilised to model the data in each bootstrap data set as previously described in this chapter.

ICH E14 [229] terms a 'Negative' thorough QTc trial to be one in which the IUT null-hypothesis, Equation (8.1), is successfully rejected. A 'Positive' study is a study in which the IUT null-hypothesis, Equation (8.1), is not successfully rejected. We will adopt this terminology for this technical appendix. The inclusion of apostrophes about these terms serves to denote this ICH-E14 based interpretation (to differentiate it from false positive and negative rates traditionally used in statistical testing relating to Type 1 and 2 error rates). This study was conducted in support of discussions during the [229] generation, and discussion of Δ of 8 msec are included for completeness.

Several properties of the IUT are of interest with regard to testing of Equation (8.1):

1. The rate at which false 'positive' studies occur for choices of 8 and 10 msec for the acceptance boundary Δ,
2. The bias present in the maximum observed mean across the post-dose sampling interval.

The identification of hazard (using the lower bound of the 90% confidence interval) is of statistical interest in this procedure (as described in the previous section), and we also tabulate findings related to this rate of occurrence.

As Table 8.8 shows, the IUT is a biased test in that at the 8 msec ($\Delta = 8$) acceptance boundary greater than 95% of the studies resulted in a 'positive' assessment of the IUT Equation 8.1. However, this bias is in favor of patient safety and is thus conservative in a regulatory sense. When the true difference in treatments is null (the assumption in most cases where Equation (8.1) will be tested), $n = 30$ to 40 subjects provides at least 85% power to successfully reject Equation (8.1). However, when the true difference between treatments is 4 msec, 46-61% of studies will result in false 'positive' findings ($n = 30$ to 40). When the true difference between treatments is 5 msec (a level of clinical concern as described in [229]), only 20-30% of data sets will be deemed 'negative' ($n = 30$ to 40).

Table 8.8 *Rate of 'Positive' [229] Findings in 2000 Bootstrap Data Sets for Assessment of Equation (8.1) with Acceptance Boundary* $\Delta = 8$ *msec*

True Mean Diff.	$n = 10$	$n = 20$	$n = 25$	$n = 30$	$n = 40$	$n = 50$	$n = 60$
0 msec	78	41	28	15	6	2	1
4 msec	92	78	70	61	46	34	25
5 msec	95	89	83	80	70	61	52
8 msec	99.6	99.3	99.3	99.2	99.4	99.2	99.5

Table 8.9 shows the IUT findings when ($\Delta = 10$ msec) in keeping with the [229] criteria.

The maximums among the 10 post-dose adjusted mean differences between treatments are positively biased in thorough QTc trials. See Table 8.10. In data sets with $n = 30$ to 40, positive bias was observed

Table 8.9 *Rate of 'Positive' [229] Findings in 2000 Bootstrap Data Sets for Assessment of Equation (8.1) with Acceptance Boundary* $\Delta = 10$ *msec*

True Mean Diff.	$n = 10$	$n = 20$	$n = 25$	$n = 30$	$n = 40$	$n = 50$	$n = 60$
0 msec	56	15	8	4	1	0	0
4 msec	77	45	32	20	11	4	2
5 msec	84	61	48	37	24	14	8
8 msec	97	94	92	91	87	84	81
10 msec	99.6	99.2	99.2	99.2	99.3	99.2	99.5

to be approximately 2 to 3 msec; however, as with the IUT findings previously described, this bias is in favour of patient safety, and is not likely to be of regulatory concern. The bias decreases with increasing sample size and was observed to be negligible (less than 1 msec) when the true mean difference between treatments was greater than 4 msec ($n \geq 30$).

Table 8.10 *Estimated Difference in Maximum Means (msec) Between Treatments [229] in 2000 Bootstrap Data Sets Relating to Assessment of Equation (8.1)*

True Mean Diff.	$n = 10$	$n = 20$	$n = 25$	$n = 30$	$n = 40$	$n = 50$	$n = 60$
0 msec	4.5	3.2	2.9	2.7	2.3	2.0	1.8
4 msec	6.3	5.4	5.2	5.0	4.8	4.7	4.6
5 msec	7.0	6.2	6.0	5.9	5.8	5.6	5.6
8 msec	9.6	9.0	8.9	8.8	8.7	8.6	8.6

To conclude, the bootstrap simulations indicate that the IUT test is quite conservative with respect to the identification of potential hazard (i.e., the lower bound of any of the 10 post-dose comparisons between treatments exceeding 0), see Table 8.11. In cases where there is truly no difference between treatments, the lower bound of at least one comparison in post-dose means should be expected to exceed 0 approximately 25% of the time. In keeping with Examples 8.1 and 8.2 however, when mild to severe QTc prolongation is present following a treatment, it is very likely to be detected by such comparisons even in studies as small as $n = 10 - 20$.

Table 8.11 *Rate of Potential Hazard Findings in 2000 Bootstrap Data Sets (Lower Bound of at Least One Post-dose Comparison Exceeds Null)*

True Mean Diff.	$n = 10$	$n = 20$	$n = 25$	$n = 30$	$n = 40$	$n = 50$	$n = 60$
0 msec	27	27	28	27	28	27	26
4 msec	49	66	71	77	85	90	95
5 msec	58	79	84	89	95	98	99
8 msec	83	97	99	99	100	100	100

8.5.2 Comparing QTc between Treatments Accounting for Correlation

Here we briefly consider an alternative to the intersection-union test which constrains the probability of a Type 1 error to precisely 5% and accounts for the correlation across time due to the repeated measures. Westfall's SimIntervals approach [449] is one such readily available procedure, and a copy of the SAS program is available in [450]. We will utilize Examples 8.1 to 8.3 to describe typical results from such testing, and the SAS code to do so may be found on the website.

Application of such a technique is useful if one wishes to ensure that autocorrelation in QTc data across the times of ECG sampling does not impact inference relative to the ICH E14 [229] intersection-union testing approach discussed in previous sections. The Westfall SimIntervals procedure uses the multivariate t-distribution among a set of finite comparisons of interest to simulate the appropriate confidence level for each time in the sampling interval (i). This ensures that the upper and lower confidence bounds provide exactly the desired level of confidence for all i for each comparison.

Application of this would be of value when one observes that proof of safety has not been provided by the intersection-union test, and one wishes to precisely evaluate when the potential for hazard exists among the i times of ECG sampling. As we will see, QTc data are so highly correlated that the Westfall adjustment does not significantly alter our inference relative to the unadjusted procedures shown in Tables 8.3 to 8.5.

Table 8.12 *Mean Changes (90% CI) Between Regimens Following a Single Dose in Example 8.1 ($n = 41$) Corrected for Correlation*

Comparison	Time	Adj. Lower 90% Bound	Adj. Upper 90% Bound
C-F	0.5	1.3018	7.6827
	1	4.9926	11.3734
	1.5	2.8216	9.2024
	2.5	0.5540	6.9348
	4	2.1040	8.4848

Estimated 90% Quartile = 2.2935 (C-F)
C = Therapeutic Dose
F = Placebo			

In Table 8.12, it is observed that significant QTc prolongation is again observed at all times of ECG sampling (as assessed by the lower bounds). Note that the confidence intervals are slightly wider that those presented in Table 8.3; however, the potential for statistically significant hazard is still present even when adjusting for the correlations in QTc over time. Findings for potential hazard relative to placebo in Regimens D and E are also similar to the unadjusted analyses of Table 8.3, and we leave inspection of such results to the reader using the available code. As inference for hazard is also not impacted by adjustment for correlation in Examples 8.2 and 8.3, we leave such an exercise to the reader.

8.5.3 Accounting for Replication in Thorough QTc Assessments

In the sections of this chapter the averages across any subsamples (r) at any time of ECG collection i were used in modelling.

Although this is common practice in the industry, it is easy to use these values themselves in the model for QTc [206] using the following code. The replicates (r) are nested within each time interval (i) in a manner chosen at the time of experimental design. Statistically, we will treat them as a fixed effect denoted `r` in the SAS code with levels $1, 2, \ldots, r$ where r is the number of such subsamples at each time of ECG assessment. The value r need not be equal across times.

```
proc mixed data=for_an method=reml
    ITDETAILS CL scoring=50
    maxiter=200;
    class subject period rel_time regimen r;
    model qtcf=qtcfb period regimen rel_time
    period*rel_time regimen*rel_time
    /DDFM=KENWARDROGER S outp=out;
    random intercept r/subject=subject;
    repeated rel_time/type=CS
    subject=subject*period;
    lsmeans rel_time*regimen/corr cov;
    ods output LSMeans=means; run;
```

Note that `r` is now included in the random statement of the model. This commands SAS to partition the observational error (associated with the replicated measurements) from the variation between-subjects (the intercept), across time (the repeated statement), and from within-subject variation in QTc (the residual).

Also note that in this code, the `type=CS` is specified instead of the auto-regressive time-series structure specified previously. The more complex the variance-covariance structure (and this one is pretty complex), the more often SAS `proc mixed` will fail to converge in REML modelling.

In the interests of parsimony, here a compound-symmetric variance-covariance structure [368] was specified for the purposes of providing an example which will converge rapidly in SAS; however, other variance structures are more appropriate to the data, and by careful specification of starting values, SAS may be made to converge. Interested readers may experiment with data to be found on the website (SAS data set, `exam_r`) for this purpose.

CHAPTER 9

Clinical Pharmacology Efficacy Studies

Introduction

In the later years of my career in clinical pharmacology, I was transferred to a strategic job on a committee which oversaw early clinical development of drugs in humans. A big part of this job was managing the interactions of the biostatistics and data management organization with a bunch of 'data-happy' clinicians. This adjective 'data-happy' refers to medics who love to collect data and want someone to analyse it until, as they say, the data pleads for mercy. Most often it is the statistician involved who ends up pleading for an end to the analysis. Data seldom speaks for themselves; someone usually has to interpret them. It is beneficial when working with such 'data-happy' people to train them to perform such exploratory statistical analyses themselves. Such an action tends to cure their state of 'data-happiness' quite effectively.

To clarify, clinical pharmacologists are paid, and in many cases earn their higher educational degree, developing new markers of clinical activity. As with QTc (described in Chapter 8), these take on the attributes of alphabet soup, in most cases, with the addition of numbers where the pharmacologists run out of letters - for example, CRP, IL8, IL5, LDL, VLDL, VLDL1, VLDL2, etc. Unlike statisticians, in general, they do not seem to have thought to introduce Greek letters, instead they just add more letters and numbers. My personal belief is that this is because the word-processing software packages they most use make it difficult to use Greek letters....

In any event, the point of measuring such markers in humans, and describing their behavior over time and relative to dose, is to detect the clinical effect of drugs on the body. This obviously is of great potential benefit. If one can measure such activity in the body, and if such activity is predictive of clinical outcomes (like stroke or myocardial infarction), then one could, in theory, predict the efficacy of drugs early in drug development! Even if it is not directly predictive, such knowledge should, in theory, allow one to improve understanding of how a drug works. Such knowledge of method of action is hoped to be beneficial.

My 'data-happy' clinicians were always excited about such endpoints, *and often wondered why I was not. They usually put it down to, 'Statisticians are just not interested in the science of such matters....' In truth, I was interested, but after many years in clinical pharmacology, I had made a conscious decision not to get excited about (or too involved in) such 'data-happy' clinical things because:*

1. *There is a lot more involved in predicting clinical outcomes than just showing that a marker is correlated to clinical outcome, and*
2. *One comes to realize that efficacy is all well and good, but safety comes first (and, as we saw in Chapter 8, is an evolving topic).*

If one cannot find a safe and well-tolerated dose range (which is what early phase development is all about), then it really does not matter how efficacious the drug is. In my experience, most drugs fail in drug development because one cannot achieve a dose that is high enough such that the drug works without untoward side-effects, not because the drug does not work.

All this said, evaluation of drug efficacy and method of action data is an important part of clinical pharmacology, and this chapter will cover some methods for modelling the behavior of such data. We first briefly review some topics related to nomenclature, assumptions, and the statistics employed for this purpose.

9.1 Background

Traditional statisticians often fail to recognize the 'learning' nature of clinical pharmacology drug development. Some have suggested that this is due to the traditional techniques inherent to how statistical science is taught at many universities. Students are taught by rote to test predetermined hypotheses using direct, confirmatory methods (like those employed in bioequivalence testing). Few assumptions about the data are made, and one in essence achieves a positive or negative outcome.

Clinical pharmacology assessments of efficacy, however, focus on learning about a compound and its properties in humans, not confirming that it has or does not have activity. In the eyes of a drug developer, a compound may be presumed to have some level of efficacy in humans. The effect may not be clinically relevant, but that is another separate issue to be determined and studied later in drug development.

First, a drug developer should learn whether the compound does roughly what one expects in humans. This approach lends itself to indirect statistical assessment (see Bayesian discussion in Chapter 5). In this chapter, we will use commonly applied traditional modelling methods and supplement them with a basic Bayesian program to illustrate the use of such procedures.

As described in Chapter 2, in drug development, one should first define a safe and well-tolerated dose range in normal healthy volunteers [225], [402] (see also Chapters 7 and 8). Traditionally, ad hoc assessment of drug activity occurs in Phase I sub-chronic dosing studies. One presumes that dose and exposure have some relationship to outcome, and applies models to quantify this expectation.

Lack of a quantified relationship in markers of human activity in Phase I is not unexpected. This can occur for many reasons, such as low sample size, lack of relevant markers of pharmacodynamic activity in a normal healthy population, or lower exposures in normal healthy volunteers relative to that to be applied in patients with the disease to be treated. As described in Chapter 7, cross-over designs (randomized or non-randomized) are traditionally applied to enhance the information gained from such trials.

Once a safe and well-tolerated dose range has been defined in Phase I studies (see Chapter 7), a pilot study is usually conducted with the new drug in a small group of patients (Phase IIa). This study is sometimes referred to as providing 'Proof-of-Concept' [402] in that it is expected to provide drug developers with some confidence in their notion that the drug will work. Again, one presumes that dose and exposure have some relationship to pharmacologic markers of drug activity (like blood pressure for example), and models are applied to quantify this relationship. Unless the disease state is markedly unstable over the length of dosing (usually limited to a month in Phase IIa due to the toxicology coverage), randomized or non-randomized cross-over designs are again the design of choice in this setting as they provide better information [399]-[400] to build the models under consideration. Sample size is limited so that if the drug proves to be unsafe in the patient population, dosing and the drug's development can be halted in a timely fashion.

ICH [225] and FDA [134] guidance on the topic calls for parallel group trials to assess such information in light of concerns with carry-over effects confounding the assessment of treatment [237]. However, such a position is logically inconsistent from a drug developer's perspective. In Phase IIa, developers work on the assumption that drug has some level of activity. Detection of a carry-over effect in a placebo-controlled cross-over trial would constitute a positive finding as the drug would have some pharmacodynamic activity to be 'carried-over'!

The information desired at this stage of development is **not** confirmatory, and those seeking a yes or no solution to questions relating to developing a drug (sometimes known as a go or no-go decision in business) are likely to be disappointed. Limited scientific evidence of clinical efficacy and understanding of method of action are generated. This information serves to modify the level of confidence a sponsor has in the likely success of a compound, hopefully (but not necessarily) in a positive manner. Often, due in part to the small sample sizes in Phase IIa, many proof of concept studies are inconclusive in a traditional statistical sense [228]. Expert judgment is usually called for in interpretation of the results.

In terms of statistics, one should establish a quantitative relationship between dose or exposure with a pharmacologic effect using a model. Many assumptions are made. For example:

1. That dose and exposure are related to the marker.
2. That the relationship between dose and exposure with the marker of interest can be expressed in a mathematical model.
3. That markers of pharmacologic activity in patients substitute for assessments of longer term clinical benefit.
4. That pharmacokinetics in plasma can predict the concentrations at the site of pharmacodynamic action.

The model(s) to be explored need not be prespecified in such exploratory, learning trials. One generally would choose to dose a limited number of patients (in the interest of their safety) with a range of doses and placebo, measure the marker of pharmacodynamic activity, and apply models to the data in a systematic, parsimonious manner to quantify the relationship of dose to the marker of interest with some degree of desired precision.

As with the study of pharmacokinetics (as we will see in Chapter 10), such models are developed over the course of drug development to help with the dosing of subsequent larger numbers of patients. Of interest is maintaining the exposure levels in a safe and well-tolerated range while achieving exposure levels of drug sufficient to treat or cure the disease condition.

Clinical pharmacologists can (and some do) overestimate the value of such exploratory data. These data are **NOT** confirmatory of efficacy in a regulatory, market-access sense [225]. As discussed in Chapter 2, regulators in general presume that a drug is not efficacious until shown otherwise. Such findings as those described above are interesting and aid regulators in determining which dose is most appropriate for initiating and treating patients [225], [134], [162]; however, a confirmatory trial is one in which the hypotheses to be tested are stated in advance and intended to provide firm, conclusive evidence of safety and efficacy [228] for a dose and dosing regimen [225]. The statistical procedures to test for success or failure of the drug to provide benefit are prespecified [228], and a determination of a positive (or negative) outcome is straightforward. Readers interested in more details around the design and analysis of confirmatory clinical trials should see [151] for an excellent summary.

The benefits of applying exploratory clinical pharmacology techniques are tangible [296], [355] improving the information gained from drug development for drug labelling and marketing while speeding study completion and subsequently (presumably) regulatory approval. Shortfalls of these procedures result from lack of education, validation, and analytic tools and procedures for their application [340], [177]. One key shortcoming, highlighted in [355], pertains to lack of knowledge of the predictive value of the pharmacodynamic marker for clinical effect (in future studies), and we now turn to further discussions on the characterization on pharmacodynamic response prior to discussing the data, models, and statistics used in such techniques.

As discussed in Chapter 2, ‘biomarkers’ or biological markers are endpoints which are ‘a physical sign or laboratory measurement that occurs in association with a pathological process and that has a putative diagnostic and/or prognostic utility’ [263], [32]. This essentially means that a biomarker is an endpoint we can measure in clinical pharmacology trials (like those described in Phases I and IIa) and presumably has something to do with the disease we are studying and are hoping will be impacted (for the better) by the drug being developed.

In contrast, surrogate markers [263], [32] are a subset of the biomarkers that can serve as a substitute [425] for a clinically meaningful endpoint. These clinical endpoints (also sometimes called outcomes) are a measure of how a patient with the disease to be treated ‘feels, functions, or survives’ [263]. Lesko and Atkinson [263] further subdivide the category of clinical endpoints into:

1. Ultimate outcome - a clinical endpoint such as survival, survival time, onset of serious morbidity, or symptomatic response that captures the benefits and risks of therapeutic intervention.
2. Intermediate endpoints - a clinical endpoint that is not the ultimate outcome but is nonetheless of real clinical benefit.

Clinical pharmacology assessments of efficacy focus mainly on biomarker assessment with some limited assessments of surrogate markers in relation to dose and concentration in plasma. Where possible, the concentration of drug at the site of action may be probed if possible. Measurement of clinical endpoints usually requires lengthy studies and significant investment. Therefore, such studies are generally not undertaken in modern drug development until sufficient confidence is generated by biomarker and surrogate marker data to reassure sponsors that the investment is worth the risk. Clinical pharmacology studies therefore do not provide direct assurance of safety and efficacy in clinical endpoints.

Establishing a biomarker as a surrogate marker is not a simple process. Temple [425] describes several criteria that must be assessed, studied, and validated before such can occur:

1. Biological plausibility including (but not limited to) consistent, extensive, and quantitative epidemiologic evidence, credible animal models, well-understood disease pathogenesis, drug mechanism of action, and surrogacy relatively late on the biological path.
2. Success in clinical trials including (but not limited to) showing that the effect on the surrogate has predicted outcomes with other drugs of the same pharmacologic class and in several other classes of drug.
3. Risk benefit and public health considerations including (but not limited to) serious or life-threatening illness with no alternative therapies.

Few endpoints would be expected to fulfill such criteria for surrogacy [147], but one thing that is very clear from the above criteria is that 'A correlate does not a surrogate make.' [147] Beware of such claims.

Some elements of defining a biomarker as a surrogate endpoint are statistical, and we refer interested readers to Prentice's classical work on the topic in [348] and an excellent comprehensive summary in [55].

For the purposes of further discussion in this chapter, we will assume that the drug of interest has been shown to impact biomarkers in animal models and that there is reason to believe dosing with drug will translate into similar pharmacodynamic effects on relevant human biomarkers. Further discussion will therefore focus on modelling of the drug, concentration, and biomarker relationship. As in previous chapters, we will focus on the application of commonly used techniques using clinical data from previous trials. We first consider data generated in Phase I sub-chronic dosing studies (see Chapter 7 for the definition and design of such trials) followed by consideration of data from patients in Phase IIa. The chapter concludes with consideration of methodology studies to elucidate method of action.

9.2 Sub-chronic Dosing

The design of sub-chronic dosing studies is described in Chapter 7. In addition to the safety assessments conducted during such trials, a plethora of data on pharmacodynamic endpoints is sometime collected to elucidate the mechanism of action of the drug being studied. These data may consist of laboratory, gene expression, or protein expression endpoints, for example.

All these pharmacodynamic data are presumably correlated with each other in one way or another. Their interrelationship may be defined in a cascade manner, in that drug treatment impacts one biological mechanism, which impacts another, which impacts another causing responses along the way. Responses may also result from parallel biological processing of drug, in that drug treatment impacts multiple mechanisms of action in parallel, e.g., one in the heart and one in the liver at roughly the same time perhaps. Several techniques are available to assess whether treatment has an effect in such large data sets [309], [311], and we will utilize one commonly used technique (see Ch. 31 of [307]) in this section to test for treatment effects over time in sub-chronic dosing trials using a data set of gene expression data before and after treatment with drug or with Placebo.

See Table 9.1. Here subjects were randomly assigned to regimen (placebo or drug treatment), had their biomarker endpoints measured on day -3, and were then dosed for 14 days with the regimen to which they were assigned with another biomarker assessment occurring on the last day of dosing (day 12).

This was a very simple sub-chronic dosing study, and looked at only one dose and placebo. In general, more doses are included in such studies allowing for more sophisticated assessment of dose-response [225]. For this data set, a simple model may be used to describe the data:

$$Y_{ijk} = \Gamma_j + \Upsilon_k + \Omega_{jk} + \Sigma_{ijk}$$

where Y_{ijk} is the matrix (data arranged in columns by endpoint) of observations for the endpoints of interest, Γ_j is a matrix which denotes day -3 or 12, Υ_k is placebo or drug treatment, and Ω_{jk} denotes day-by-regimen interaction (see Chapter 3), with Σ_{ijk} denoting residual variability. Here we are interested in the significance of the Ω_{jk} as this would signal that the regimens are behaving differently across time in some manner for at least one endpoint.

The element of Ω_{jk} (and the other matrices) are arranged to correspond to the endpoint to which they relate. For example, for the data in Table 9.1,

$$\Omega_{jk} =$$
$$\begin{bmatrix} \omega_{L11} & \omega_{L12} & \omega_{L21} & \omega_{L22} \\ \omega_{M11} & \omega_{M12} & \omega_{M21} & \omega_{M22} \\ \omega_{N11} & \omega_{N12} & \omega_{N21} & \omega_{N22} \\ \omega_{O11} & \omega_{O12} & \omega_{O21} & \omega_{O22} \\ \omega_{P11} & \omega_{P12} & \omega_{P21} & \omega_{P22} \end{bmatrix}$$

for endpoint L, M, N, O, and P in each row, respectively, where, for example, ω_{L11} denotes the mean effect of treatment with drug on day -3 and ω_{L12} denotes the mean effect of treatment with drug on day 12, and so on.

We are interested in testing the null hypothesis:

$$\omega_{L11} = \omega_{L12} = \omega_{L21} = \omega_{L22},$$

$$\omega_{M11} = \omega_{M12} = \omega_{M21} = \omega_{M22},$$

$$\omega_{N11} = \omega_{N12} = \omega_{N21} = \omega_{N22},$$

Table 9.1 *Example 9.2.1: Pharmacodynamic Biomarker Data from an Exploratory Sub-chronic Dosing Study*

Subject	Day	Regimen	L	M	N	O	P
102	-3	P	7.67	7.95	9.47	10.17	4.65
102	12	P	6.77	7.50	8.89	9.60	4.47
103	-3	D	7.60	7.69	9.60	9.41	5.04
103	12	D	7.33	7.91	9.39	9.39	5.11
104	-3	D	7.61	7.58	9.60	8.93	5.32
104	12	D	7.36	8.02	9.86	9.70	5.46
201	-3	P	6.00	7.24	8.56	8.99	3.87
201	12	P	6.66	7.52	8.88	9.60	4.38
202	-3	D	8.04	8.35	9.46	9.75	5.26
202	12	D	7.32	7.74	9.36	9.10	4.85
204	-3	D	6.83	7.53	8.75	8.82	4.67
204	12	D	6.79	7.54	8.75	8.74	4.84
205	-3	P	7.33	7.81	9.27	9.52	5.15
205	12	P	7.06	7.49	8.98	8.96	4.32
208	-3	P	7.36	7.71	9.29	9.81	5.23
208	12	P	7.43	7.86	9.48	9.32	5.23
209	-3	D	7.60	7.83	9.70	9.68	5.23
209	12	D	6.70	7.64	8.90	9.07	4.44
210	-3	D	6.76	7.69	8.86	9.05	5.12
210	12	D	6.65	7.66	8.61	9.32	4.91
211	-3	P	7.15	7.91	9.73	10.22	5.26
211	12	P	7.29	7.98	9.20	10.17	5.11
213	-3	P	6.76	7.82	9.41	9.44	5.03
213	12	P	7.50	7.95	9.37	9.38	5.18

P=Placebo; D=Dose of Drug for 14 Days

$$\omega_{O11} = \omega_{O12} = \omega_{O21} = \omega_{O22},$$

$$\omega_{P11} = \omega_{P12} = \omega_{P21} = \omega_{P22}$$

versus the alternative that at least one of the ω_{jk} differs from the others for at least one of the endpoints.

Computation of estimates for the elements of the various matrices (like Ω_{jk}) is a complex topic beyond the scope of this book. See [309] and [311] for a description. SAS automatically performs some of these calculations [368] using a procedure similar to `proc mixed` known as `proc glm`. Code to so for this study is as follows.

Sub-chronic Exploratory Multivariate Data Analysis 9.2.1 - SAS `proc glm` *code:*

```
proc glm data=my.sc1a;
    class day regimen;
    model L M N O P = day
    regimen regimen*day;
    MANOVA h=day regimen regimen*day;
    run;
```

Here `proc glm` is called and directed to assess data from the permanent data set `my.sc1a` (available on the website accompanying this book). The `class` statement again specifies the descriptive variables of the model, and the `model` statement specifies that endpoints L, M, N, O, and P (a subset of those available, included here for simplicity) be modelled as a function of day, regimen, and day-by-regimen interaction. The `MANOVA` statement specifies that SAS should construct tests to assess whether the study days are different (pooling across regimens), whether the regimens are different (pooling across study days), and (most of interest) whether response to treatment is different between regimens over time, testing the null hypothesis described above for these endpoints simultaneously.

Selected SAS outputs appear as follows:

Sub-chronic Exploratory Multivariate Data Analysis 9.2.1 - Selected SAS `proc glm` *output:*

```
MANOVA Test Criteria and Exact F Statistics for
 the Hypothesis of No Overall DAY*REGIMEN Effect
  H = Type III SSCP Matrix for DAY*REGIMEN
  E = Error SSCP Matrix
Statistic         F Value    Num DF   Den DF   Pr > F
  Wilks'Lambda    0.94          5        16    0.4801
```

Based on the p-value for day-by-regimen interaction ($p = 0.4801$), there is very little evidence to suggest that treatment with drug impacts the biomarkers considered here (endpoints L, M, N, O, and P) over the course of 14 days of treatment. This is not unexpected in Phase I drug development as described in the previous section. At worst, such data are valuable in that they provide variability estimates for use in better sizing subsequent trials. At best, one may see some evidence of treatment effects that would also aid in designing more definitive trials later in drug development.

The downside of utilization of a multivariate statistical procedure as described above is that it is known [307] to be less powerful (prone to false negatives) than univariate methods (which we will now discuss). However, such an approach serves as a handy tool for rapid assessment of whether there is value in extensive data mining of a large pharmacodynamic data set. Interested readers may use the data from other biomarker endpoints contained in `my.sc1a` and another data set, `my.sc1b`, on the website that accompanies this book, to explore such multivariate methods and alternative statistical procedures.

A more powerful approach, a dose-response analysis, will now be discussed. Here, low density lipoprotein, LDL, was measured (decreasing this surrogate marker results in clinical benefit [101]), before and after sub-chronic treatment with a randomly assigned dose of drug or placebo in each normal healthy volunteer subject. The objective of modelling in this case was to assess whether there was evidence of a response to dose in this population (normal healthy volunteers).

Table 9.2: Example 9.2.2: Dose, Pharmacokinetic, and Low Density Lipoprotein Data from a Sub-chronic Dosing Study

Subject	Dose			Baseline	Post-Trt
		AUC	Cmax	LDL	LDL
56	0	0.00	0.000	2.18	2.22
63	0	0.00	0.000	3.53	4.47
67	0	0.00	0.000	2.85	3.01
73	0	0.00	0.000	1.37	1.74
74	0	0.00	0.000	2.71	2.26
86	0	0.00	0.000	2.93	2.78
87	0	0.00	0.000	2.80	3.09
91	0	0.00	0.000	2.40	2.59
94	0	0.00	0.000	5.33	5.36
100	0	0.00	0.000	2.04	2.32
103	0	0.00	0.000	3.31	3.21
112	0	0.00	0.000	1.92	2.05
47	5	5.11	0.423	3.03	2.89
48	5	8.13	0.620	2.59	1.95
49	5	8.01	0.627	2.05	1.72
50	5	6.67	0.480	3.06	2.66
52	5	7.38	0.591	4.01	2.80
53	5	5.17	0.390	3.27	3.52
54	5	8.16	0.569	3.25	3.35
55	5	6.23	0.483	2.52	2.38
AUC and Cmax assumed 0 if Dose was 0					

Table 9.2: Example 9.2.2: Dose, Pharmacokinetic, and Low Density Lipoprotein Data from a Sub-chronic Dosing Study

Subject	Dose			Baseline	Post-Trt
		AUC	Cmax	LDL	LDL
57	5	3.36	0.316	2.14	2.14
60	10	11.22	0.962	3.98	3.13
61	10	8.21	0.723	1.70	1.78
62	10	20.85	1.861	2.96	2.05
64	10	16.48	1.169	2.28	2.54
65	10	6.79	0.574	3.09	3.64
66	10	18.08	1.303	2.13	1.77
69	10	10.51	0.883	2.15	1.78
71	10	13.97	1.056	3.45	2.98
72	10	13.80	1.157	2.77	2.25
95	20	30.35	2.220	2.47	1.88
99	20	53.11	3.902	2.31	1.88
102	20	38.61	2.517	3.13	2.93
104	20	29.33	2.219	3.68	4.27
105	20	26.20	1.844	3.20	3.10
106	20	29.47	1.893	3.16	3.40
107	20	27.55	1.965	3.35	3.18
108	20	19.97	1.447	1.84	1.98
110	20	35.91	2.322	3.44	3.36
AUC and Cmax assumed 0 if Dose was 0					

Previous experience indicated that LDL was log-normally distributed in normal healthy volunteers, so in a manner similar to pharmacokinetic analysis this endpoint was log-transformed for analysis following correction for baseline (in this case, simply by taking the ratio of posttreatment LDL to baseline LDL). Only a limited response was expected in normal healthy volunteers, and for the purposes of this example, a power model (see Chapter 7) was utilised of the form:

$$y_k = \alpha + \beta(ld) + \varepsilon_k$$

where β is the slope parameter of interest regressed on logDose (parameter ld) for the log-transformed ratio of posttreatment LDL to baseline LDL for each subject k. Note that we do not have repeated measurements within a subject, so there is no term denoting each subject as their own control nor denoting the repeated measures. This is often the case in Phase I designs as such pharmacodynamic assessment are (relatively) expensive and are of limited value due to the normal healthy population being studied and the expected portfolio attrition rates.

In this case, the power model was selected for use as normal healthy volunteers in general do not have high LDL values, and therefore may be expected to show only a limited response to treatment (if at all). To include the placebo data (null dose) in the power model, the dose needs to be set to a value greater than 0 prior to log-transform such as 0.000001 using SAS statements such as those following in a data step (see also `dose response.sas` on the website accompanying this book):

```
if dose=0 then dose=0.000001;
```

In general, one would build such a nonlinear dose-response model after first investigating the fit of a linear dose response model [314], [179], [384], [307]. In this case, the fit of a linear model is poor and indicative of heterogeneous variance. We leave confirmation of this point to the reader and encourage readers interested in more details of model building to investigate the above references. SAS code for this analysis is:

Sub-chronic Dose Response Data Analysis 9.2.2 - SAS `proc mixed` *code:*

```
title 'Log-Ratio Power Model';run;
proc mixed data=sc2a;
    class subject;
    model ldlchg=lndose/s cl ddfm=kenwardroger
    outp=out outpm=predmean;
    run;

proc print noobs data=predmean;
    where subject<10;
    var dose alpha pred lower upper;
    run;
```

Here the log-transformed ratio of posttreatment to baseline LDL is fitted versus logDose. Residuals are output to SAS data set `out` for use in assessing model fit (not shown), and predicted mean values are output to a SAS data set `predmean`. To derive estimates of dose response one includes 'dummy' subjects (in this case, subjects 1 to 9 corresponding to doses of 0 to 80 mg, see `dose response.sas`) with dosing information in the analysis data set but with no data on LDL response. As no LDL data are available for these 'dummy' subjects, they are not included in model fitting, but SAS provides estimates of effect for these subjects in the `predmean` data set. Selected SAS output is as follows.

Sub-chronic Dose Response Data Analysis 9.2.2 - Selected SAS `proc mixed` *output:*

```
The Mixed Procedure
Solution for Fixed Effects

 Effect      Estimate   Pr > |t|
 Intercept   -0.06567   0.0104
 lndose      -0.00854   0.0086

      Type 3 Tests of Fixed Effects
                Num     Den
Effect          DF      DF    F Value    Pr > F
lndose           1      37       7.71    0.0086

DOSE    Alpha       Pred        Lower        Upper
 0      0.05     0.05226     -0.03131      0.13584
 5      0.05    -0.07941     -0.13290     -0.02592
10      0.05    -0.08533     -0.14109     -0.02957
20      0.05    -0.09125     -0.14951     -0.03299
30      0.05    -0.09471     -0.15453     -0.03489
40      0.05    -0.09716     -0.15813     -0.03620
60      0.05    -0.10062     -0.16326     -0.03799
80      0.05    -0.10308     -0.16693     -0.03923
```

In this case, we observed that LDL (adjusted for baseline) decreases with increasing logDose (estimate of -0.00854 for β with $p = 0.0086$). The values `Pred`, `Lower`, and `Upper` may be exponentiated to estimate dose-response in LDL (adjusted for baseline) on the original scale resulting in the findings of Table 9.3.

Here we observe that dosing in normal healthy volunteers resulted in decreases of 8 to 10% in LDL (adjusted for baseline). This is promising data (in terms of effect on a surrogate marker in Phase I). However, overinterpretation of such data is not recommended. Data from normal healthy volunteers can only predict clinical outcomes under carefully controlled circumstances (see Chapter 8 for an example).

Here, these findings should increase confidence in the drug's potential to be useful in the clinic, but such data are not definitive (as patients with disease have not yet been assessed). Pairwise testing between mean responses (for example direct comparison of 5, 10, and 20 mg to placebo) as is often done in Phase II-III dose-response testing [254] is not recommended here as such analyses are typically misleading in trials of such limited sample size.

Note that sample size selection is driven by safety considerations in

Table 9.3 *LDL Dose-Response (Ratio relative to Baseline LDL) with 95% Confidence Intervals in Sub-chronic Dosing Study Example 9.2.2*

Dose	Estimated Effect	95% CI
0	1.05	0.97-1.15
5	0.92	0.88-0.97
10	0.92	0.87-0.97
20	0.91	0.86-0.97
40	0.91	0.85-0.96
60	0.90	0.85-0.96
80	0.90	0.85-0.96

such designs (see Chapter 7). Precision in statistical study results may be evaluated at the design stage using techniques similar to those described in Chapter 10 and is not discussed further here.

Models such as the above may also be useful to provide a preliminary check for association between pharmacokinetic and pharmacodynamic responses. In this case, steady state AUC and Cmax did not appear to be related to baseline-adjusted LDL changes ($p = 0.8026$ and $p = 0.6549$ for logAUC and logCmax, respectively). Absence of a significant relationship may not preclude that such an association exists [238]. As described previously, such an observation may occur due to low sample size and may be related to not having a model accounting for all relevant biologic information. In this particular case, for example, it was thought that the drug worked in the liver such that plasma pharmacokinetics were not predictive of concentrations at the site of action. Plasma concentrations in the liver may be modelled by extending the findings of Chapter 10 to allow for another compartment. As shown in Chapters 7 and 10, dose and AUC are to some extent confounded and their use in a model simultaneously is therefore of questionable validity [354], potentially leading to model over-specification [314].

We now turn to modelling and interpretation of data in clinical pharmacology studies of patient populations.

9.3 Phase IIa and the Proof of Concept

For purposes of illustration, assume that a proof-of-concept trial was desired to test whether the LDL response in normal healthy volunteers described in the last section would result in clinical benefit when given to patients. LDL was again to be used as a surrogate marker of clinical benefit for the purposes of this trial.

Recall that about a 10% decrease LDL was observed in normal healthy volunteers. When dosing in patients, it might be expected that approximately twice this magnitude would be observed as:

1. Patients with disease would be recruited with higher LDL (than the subjects in Phase I) allowing more of an effect of drug to be observed,
2. Dosing in patients was planned to be of at least twice the duration of Phase I, and
3. Animal efficacy data indicated that drug would be more effective than observed in the Phase I sub-chronic dosing study.

The proof-of-concept Phase IIa trial was designed in a standard [228] pessimistic fashion under the assumption (null hypothesis) that drug would have no effect on LDL. The alternative to be tested was that treatment with drug would result in a 20% decrease (accounting for baseline) relative to placebo.

Patients with high LDL (who are not already taking some form of medication) are not easy to find. This resulted in a lengthy trial duration to recruit only 15 patients in a 2 × 2 cross-over design. LDL data from this trial may be found in Table 9.4. After a baseline LDL assessment in each session, patients were dosed with a drug expected to lower LDL level (or placebo) for 6 weeks.

Table 9.4: Example 9.3.1: Low Density Lipoprotein Data from a Proof-of-Concept Study

Subject	Sequence	Per	Reg	Post-Trt LDL	Baseline LDL	Analysis Endpoint
2472	PA	1	P	101	98	0.031
2472	PA	2	A	89	110	-0.212
2530	AP	1	A	140	159	-0.132
2530	AP	2	P	146	151	-0.034
2535	PA	1	P	100	86	0.150
2535	PA	2	A	106	82	0.257
2540	PA	1	P	163	135	0.182
2540	PA	2	A	139	143	-0.027
2544	PA	1	P	160	147	0.086
2544	PA	2	A	99	123	-0.220
2546	AP	1	A	85	103	-0.186
2546	AP	2	P	81	92	-0.126
A=Drug Treatment; P=Placebo Endpoint=Natural-log of Post-Trt LDL to Baseline LDL						

Table 9.4: Example 9.3.1: Low Density Lipoprotein Data from a Proof-of-Concept Study

Subject	Sequence	Per	Reg	Post-Trt LDL	Baseline LDL	Analysis Endpoint
2548	AP	1	A	106	115	-0.077
2548	AP	2	P	96	111	-0.139
2549	PA	1	P	125	142	-0.128
2549	PA	2	A	116	126	-0.083
2560	PA	1	P	155	178	-0.140
2560	PA	2	A	108	151	-0.331
2562	PA	1	P	104	124	-0.170
2562	PA	2	A	97	104	-0.077
2650	AP	1	A	128	139	-0.087
2650	AP	2	P	132	124	0.061
2659	PA	1	P	120	108	0.102
2659	PA	2	A	101	116	-0.143
2668	PA	1	P	151	128	0.167
2668	PA	2	A	128	163	-0.241
2712	AP	1	A	120	124	-0.032
2712	AP	2	P	108	139	-0.251
2755	PA	1	P	132	147	-0.111
2755	PA	2	A	132	151	-0.137

A=Drug Treatment; P=Placebo
Endpoint=Natural-log of Post-Trt LDL to Baseline LDL

The SAS code used to analyse the analysis endpoint of Table 9.4 is the same as that used to analyse 2 × 2 cross-over studies in Chapter 3, and is not reproduced here. Readers interested in the code may find it on the website accompanying this book.

The findings ($n = 15$) indicated that treatment with drug lowered LDL by only approximately 7% relative to placebo (the effect size on the log-scale was -0.07177 with 90% confidence interval of -0.1544 to 0.01090). As the upper bound of the 90% confidence interval exceeded zero, the null hypothesis (that drug does not significantly change LDL relative to placebo) was not rejected. This would therefore be regarded as a 'failed' study.

However, the findings provide some useful information [187], [245]:

1. The bulk of the confidence interval falls to the left of null; therefore, while we cannot conclude that this dose of drug is effective, it suggests the potential for increased doses of drug to provide significant benefit.
2. The maximum expected mean effect of this dose of drug is a 14% decrease in LDL (corresponding to the exponentiated lower confidence

limit) with the effect size most likely falling around 7%. Such a small decrease might be desirable (and clinically relevant) in some patient population.

Thus while failing to reject the null hypotheses, the study has provided some degree of useful information.

The above approach is a traditional one, and should it be successful (as this example was not), it clearly increases confidence that the drug will be efficacious even against a pessimistic level of opinion concerning the drug's merits. Such studies need not be designed to provide such a yes-or-no answer however. Moreover, planning a traditional hypothesis testing approach, like that described here, requires a long time. One would probably wait to analyse the data until the full ($n = 15$) complement of patients complete the study.

A Bayesian analysis (described in Chapter 5) provides a ready alternative to the traditional analysis described above. Here, we may take explicit account that an effect size of approximately 10% is our expectation and express it as a prior distribution for `delta` (the effect size of treatment with drug relative to placebo). WINBUGS code to perform a Bayesian analysis on the first eight patients (approximately halfway through the study) is provided on the website accompanying this book and is the same as that utilised for bioequivalence testing in Chapter 5.

With data from only eight patients, the Bayesian analysis (see Table 9.5) provides the following expectations regarding the effect size on the log-transformed scale and original scale.

Table 9.5 *LDL Effect Size (Ratio relative to Baseline LDL) from a Bayesian Statistical Analysis of a Proof-of-Concept Study Example 9.3.1 (n = 8)*

Trt:Pbo Baseline Adj Effect Size	2.5 QTL	5 QTL	Median	90 QTL	97.5 QTL
ln-Scale Original-Scale	-.2141 0.8073	-0.1878 0.8288	-0.07697 0.9259	0.006284 1.006	0.05994 1.062
QTL=Quartile of the Bayesian Posterior Distribution					

From this Bayesian analysis, (based on the 90th percentile) we can conclude (with only $n = 8$) that the drug has approximately a 90% probability of reducing LDL relative to placebo. Conversely, there is a lesser chance (approximately 10%) that the drug treatment is the same or worse than placebo. The effect size with this dose of drug is unlikely (less than a 5% chance, based on the 5th percentile) to be greater than a 17% decrease and is most unlikely (less than a 2.5% chance, based on the 2.5th percentile) to reach the desired decrease level of 20% in posttreatment LDL relative to placebo.

If one looks closely, this is about the same amount of information one could glean from the traditional analysis and design described above, except that this Bayesian analysis approach, if used, only takes half as many patients and half the time as the original study. Bayesian design and analysis plans such as these can be very useful tools to increase a sponsor's confidence in the properties of a compound without requiring long resource-intensive studies. Such an approach is useful for internal decision making; however, use in a regulatory setting when wishing to make a claim about the properties of a drug (for the reasons discussed in Chapter 2) is of questionable validity.

An unstated reason why one often does not utilize such an approach to design and analysis is the wish to publish data from such studies. Akin to the approach to data interpretation taken by regulators, most scientific journals would question the application of such a Bayesian approach closely as such techniques are only now becoming widely used and have been the matter of some historical debate. A group-sequential approach (described in Chapter 5; and code for the interim and final analyses may be found on the website) may be used if a journal-acceptable approach is desired. Here, interested readers will observe that such a group-sequential approach provides approximately the same information as the Bayesian analysis.

We now turn to consideration of extensions of dose-response modelling involving pharmacokinetic-pharmacodynamic modelling [403]. With the publication of [225] and [134], applications of such techniques are becoming more frequent in drug development. Typically, what is done is to develop nonlinear mixed effect models [274] for pharmacokinetics in an effect compartment ([403], a hypothesized part of the body where pharmacodynamic effect is thought to be induced by drug treatment) and then relate that to a model of pharmacodynamic activity using a statistical model [286].

Specialized software is generally needed for such an activity. Several packages are described in [407]. See also [365] for a review of some data comparisons between available software packages. Many software packages are available - see http://www.boomer.org/pkin/soft.html for a list. For the purposes of illustrating the principles involved, we will make use of a data set involving dose, pharmacokinetics, and some data on QTc (see Chapter 8) using SAS from a longitudinal, repeated-measures proof-of-concept study. Other software programs may also be used to model this data (e.g., SPLUS, PKBUGS, NONMEM, NONLINMIX), and we invite interested readers to make use of the data available on the website accompanying this book to do so.

The data used in the following examples for PK-PD modelling are quite extensive. Measurement of QTc was taken over a period of eight days to assess the properties of the compound under study (utilizing doses up to 120 mg) with pharmacokinetic assessment to measure plasma concentration taken at regular intervals. See Tables 9.6 and 9.7. The full data sets may be found on the website accompanying this book.

Table 9.6: Example 9.3.2: QTc Data from One Subject in a Proof-of-Concept Study

Subject	Dose(mg)	Day	Time(h)	QTc(msec)
1	80	1	0	393
1	80	1	0.5	394
1	80	1	1	399
1	80	1	1.5	400
1	80	1	2	416
1	80	1	3	418
1	80	1	4	396
1	80	1	6	402
1	80	1	8	405
1	80	1	10	393
1	80	1	12	390
1	80	1	18	388
1	80	2	0	406
1	80	2	3	413
1	80	2	12	386
1	80	2	15	421
1	80	3	0	421
1	80	3	3	425
1	80	3	12	394
1	80	3	15	420
1	80	4	0	427
1	80	4	3	430
1	80	4	12	384
1	80	4	15	417
1	80	5	0	425
1	80	5	3	435
1	80	5	12	398
1	80	5	15	415
1	80	6	0	409

Table 9.6: Example 9.3.2: QTc Data from One Subject in a Proof-of-Concept Study

Subject	Dose(mg)	Day	Time(h)	QTc(msec)
1	80	6	3	434
1	80	6	12	388
1	80	6	15	418
1	80	7	0	420
1	80	7	3	409
1	80	7	12	398
1	80	7	15	410
1	80	8	0	407
1	80	8	0.5	411
1	80	8	1	432
1	80	8	1.5	443
1	80	8	2	455
1	80	8	3	460
1	80	8	4	428
1	80	8	6	419
1	80	8	8	382
1	80	8	10	404
1	80	8	12	388
1	80	8	18	384
1	80	8	24	409
1	80	8	36	384
1	80	8	48	388

Table 9.7: Example 9.3.2: Plasma Pharmacokinetic-Pharmacodynamic Data from One Subject in a Proof-of-Concept Study

Subject	Day	Time(h)	QTc(msec)	Conc.(ng/mL)
1	1	0	393	.
1	1	0.5	394	.
1	1	1	399	8.15
1	1	1.5	400	7.89
1	1	2	416	7.56
1	1	3	418	5.43
1	1	4	396	3.58
1	1	6	402	.
1	1	8	405	.
1	1	10	393	.
1	1	12	390	.

Table 9.7: Example 9.3.2: Plasma Pharmacokinetic-Pharmacodynamic Data from One Subject in a Proof-of-Concept Study

Subject	Day	Time(h)	QTc(msec)	Conc.(ng/mL)
1	1	18	388	.
1	1	24	.	.
1	8	0	407	2.53
1	8	0.5	411	5.26
1	8	1	432	13.9
1	8	1.5	443	14.72
1	8	2	455	17.12
1	8	3	460	12.81
1	8	4	428	9.39
1	8	6	419	5.83
1	8	8	382	3.09
1	8	10	404	.
1	8	12	388	.
1	8	18	384	.
1	8	24	409	.
1	8	36	384	.
1	8	48	388	.

We will first build a dose-response model for these data and then will supplement it with a discussion of how to build a PK-PD model for the data to illustrate the concepts involved.

Readers familiar with Chapters 7 and 8 will recognize the data in Table 9.6 as being of the general form consistent with repeated-measures data. As such, it can be modelled simply using a model of the form:

$$y_{ijk} = \alpha + \phi_j + \tau_k + (interactions) + \beta_1(dose) + \varepsilon_{ijk},$$

where α is the common-intercept, ϕ_j adjusts for study day j, τ_k adjusts for each time k, β_1 denotes the slope of dose-response. The terms of the error-term ε_{ijk} are constructed recognizing that QTc responses (y_{ijk}) are correlated across time within each day for each subject (i). The interactions (not described here) are combinations of the dose, day, and time information to study whether response to a dose of drug is dependent on the day and time of sampling. In SAS such a model can be implemented in `proc mixed` as:

Dose-Response Repeated Measures Data Analysis of Example 9.3.2 - SAS `proc mixed` *code:*

```
proc mixed data=my.poc2 method=reml;
    class subject day time;
    model qtc=dose day time
    day*time dose*day dose*time
    /DDFM=KENWARDROGER outp=out;
    repeated time/type=AR(1) subject=subject*day;
    lsmeans day*time/at dose=0 CL ALPHA=0.1;
    lsmeans day*time/at dose=25 CL ALPHA=0.1;
    lsmeans day*time/at dose=80 CL ALPHA=0.1;
    lsmeans day*time/at dose=120 CL ALPHA=0.1;
    lsmeans day*time/at dose=200 CL ALPHA=0.1;
    ods output LSMeans=my.lsmeans;
    run;
```

This model indicates (outputs not shown) that a significant, linear dose-response relationship was observed for QTc ($p < 0.0001$) and that the response changes over the course of eight days ($p < 0.0001$) and over times of ECG sampling ($p < 0.0001$). The `lsmeans` statements output the expected responses at various doses to a data set called `my.lsmeans` for further assessment, and the data set `out` may be used to assess model fit as described in previous chapters. Here the model fit as assessed by residuals appeared adequate, and Table 9.8 gives the expected responses on placebo (dose of 0 mg) on day 1 and day 8 for example.

Table 9.8: QTc Response on Placebo on Days 1 and 8 in a Proof-of-Concept Study from Modelling of Dose-QTc data in Example 9.3.2

Day	Time(h)	Dose	Mean QTc	95% CI
1	0	0	398	(390,405)
1	0.5	0	396	(388,405)
1	1	0	392	(384,400)
1	1.5	0	398	(390,406)
1	2	0	397	(389,405)
1	3	0	400	(392,407)
1	4	0	398	(390,406)
1	6	0	395	(387,403)
1	8	0	399	(391,407)
1	10	0	398	(390,407)
1	12	0	401	(394,409)
1	18	0	409	(401,418)
8	0	0	394	(386,402)

Table 9.8: QTc Response on Placebo on Days 1 and 8 in a Proof-of-Concept Study from Modelling of Dose-QTc data in Example 9.3.2

Day	Time(h)	Dose	Mean QTc	95% CI
8	0.5	0	391	(382,399)
8	1	0	398	(389,406)
8	1.5	0	398	(390,406)
8	2	0	396	(388,405)
8	3	0	395	(388,403)
8	4	0	394	(386,402)
8	6	0	392	(383,400)
8	8	0	386	(378,395)
8	10	0	390	(381,398)
8	12	0	392	(384,400)
8	18	0	394	(385,402)
8	24	0	391	(382,400)
8	36	0	389	(380,398)
8	48	0	393	(384,403)

In addition to confirming that a dose-response is evident, providing overall positive evidence of efficacy for the compound [225], (though we do not yet know which dose is best in terms of safety), the importance of the model's findings in terms of response on placebo are very important. These will figure prominently as we develop models for concentration to QTc response relationships. For the purposes of this example, we neglect the development of a pharmacokinetic compartment model. Readers interested in doing so should see Chapters 7 and 10 for more details. In this example, plasma concentration is therefore assumed to be the effect compartment where pharmacodynamic effect is caused by drug action.

The first step taken in modelling such data (see Chapter 4 of [41]) is to assess whether a linear relationship exists between concentration and response. This can easily be accommodated using the above SAS code (replacing dose with concentration). Eliminating nonsignificant terms, we use the following SAS model to examine the relationship of concentration to QTc where the term `pt_` denotes subject and the term `pkp_c` is concentration.

Concentration Response Repeated Measures Data Analysis - SAS `proc mixed` *code:*

```
proc mixed data=poc2pkpd method=reml;
    class pt_ day time;
    model qtc=pkp_c time
    /DDFM=KENWARDROGER outp=out S;
    repeated time/type=AR(1) subject=pt_*day;
    run;
```

Model fit may again be examined using the data set `out` and was observed to be adequate (not shown). Concentration was a significant ($p < 0.0001$) linear predictor of QTc with a slope of 0.38. This indicates that as drug concentration in blood increases, so too does QTc.

If the fit was not adequate, any number of other potential nonlinear models may be fitted [314]. However, by far the favorite model used in PK-PD research is the Emax model (named for one of the parameters used in the model). Boxtel et al. [41] described these models in great detail, and we shall dwell only on simple examination of Day 8 QTc and concentration data using such a model. Interested readers may apply other models using the data on the website and may find Chapter 15 of [41] helpful for additional background materials on PK-PD modelling in cardiac repolarisation.

The Emax model is described as [41]:

$$E = \frac{Emax(C)}{EC50 + C} + E_0,$$

where E is the effect being modelled, E_0 is the effect observed without any drug present, C is the concentration of drug in the effect compartment, $EC50$ is the concentration needed to cause a 50% response, and $Emax$ is the maximum effect that can occur with drug treatment. This is a nonlinear (in concentration) additive model. If concentration is not related to effect, Emax and EC50 would be zero.

Here, we are interested in assessing the following model:

$$QTc_{ij}(Effect) = \frac{Emax_i * C}{EC50_i + C} + E_0 + \varepsilon_{ij},$$

on Day 8, where the subscript i denotes subject, j denotes time, and ε_{ij} is the usual term for residual error. Such a model is easily implemented in `proc nlmixed` in SAS as:

Emax Concentration Response Data Analysis Example 9.3.2 - SAS `proc nlmixed` *code:*

```
proc nlmixed data=pkpd2;
    parms beta1=4.6 beta2=5.57 s2b1=1
    s2b2=1 s2=400;
    emax = exp(beta1+b1);
    ec50 = exp(beta2+b2);
    pred=((emax*pkp_c)/(pkp_c+ec50))+e0;
    model qtc ~ normal(pred,s2);
    random b1 b2 ~ normal([0,0],[s2b1,0,
    s2b2]) subject=pt_;
    predict pred out=pred;
    run;

*Model fit assessment;
data pred;set pred;
    st_resid=(qtc-Pred)/StdErrPred;
    run;
proc rank data=pred normal=blom out=nscore;
    var st_resid;
    ranks nscore;
data nscore;
  set nscore;
  label nscore="Normal Score";
  label stres="Residual";
  label pred="Predicted Value";
    run;
proc plot vpercent=50 data=nscore;
    plot st_resid*pred/vref=0;
    plot st_resid*nscore;
    run;
```

Here `proc nlmixed` is called and applied to a data set denoted as `pkpd2` where the placebo modelling results of the dose-response model (Table 9.8) have been used to describe E_0. In general, it is more desirable for each subject to provide such an assessment so that more informative models may be fitted [399], but such is obviously not possible with this data as subjects were not crossed over to Placebo. As in the nonlinear mixed effect examples of Chapters 7 and 10, starting values must be specified for the parameters of interest (`beta1` and `beta2`, their variances, and the residual variance). As both Emax and EC50 must be positive, the exponential function is used to allow their estimated values to be such and to accommodate subject-specific adjustment, as appropriate to the data, for each parameter. The predicted values `pred` are output as in Chapter 7 for assessment of model fit using residual plots using the above code (results not shown) which appeared adequate.

The model indicated that both Emax ($p < 0.0001$) and EC50 ($p < 0.0001$) were important in describing the QTc response. The estimates of Emax and ED50 were 86.8 (95% CI of 74.7-101) and 26.7 (95% CI of 12.6-56.2), respectively.

One should be careful with the interpretation of such a model in early phase trials. If we assume a basal QTc of approximately 400 msec in keeping with Table 9.8, one might be tempted to interpret this model as indicative that the maximum prolongation in QTc possible with this drug would be approximately 500 msec by looking at the magnitude of the upper bound of Emax. However, Emax is in this case design dependent. Dosing was terminated at the 120 mg dose in this study as prolongation was approaching a QTc of 500 msec (known, see Chapter 8, to be a level associated with a potentially fatal cardiac arrythmia). Note that the linear model predicts no such plateau in effect. Models such as these should be interpreted in tandem and developed further as drug development progresses from Phase II to file and beyond.

9.4 Methodology Studies

Methodology studies are conducted for a variety of reasons. For example:

1. To develop a new biomarker assay (an assay is a technique for measuring a biological effect in this setting),
2. To validate a biomarker's assay (to confirm that a technique is useful in practice for measuring an effect), or
3. To measure whether a drug impacts a biomarker measured by an assay.

Assay development and validation are neglected, but essential, parts of drug development, and we will not attempt to correct that situation here. Developing a new assay is primarily a matter for subject-area scientists. Statistics are used in its validation, to quantify, for example, an assay's (see Chapter 12 of [12] and [226] for more details and definitions):

1. Limits of detection and quantification,
2. Sensitivity,
3. Selectivity,
4. Accuracy,
5. Precision,
6. Reproducibility, and
7. Repeatability.

For the purposes of this section, we will assume that a validated assay has been produced and is to be utilised in the assessment of whether a drug has an effect on some biomarker. Often, given time constraints in drug development, assay development, validation, and measurement of whether a new drug has an effect are combined into one trial. We recommend against this as trials designed for such multiple purposes, under reasonably pessimistic assumptions, often are doomed to fail before they start due to overwhelming complexity.

Pessimism is usually warranted about new biomarkers and their assays. As indicated, novelty is how science and medicine develop. It is an art; however, the practical utility of such approaches in drug development takes time, experience, and accumulation of data and opinion. This does not happen overnight.

In the following example, we consider a new assay being applied to measure a new biomarker. The purpose of the trial was to assess whether drug treatment changed a biomarker thought to be related to the disease under study when assessed using a drug known (i.e., marketed) to improve the disease state. Such a study would be utilized to confirm that the known efficacious drug treatment significantly impacts a biomarker following treatment. If successful, in subsequent trials, new drugs for the same disease might be studied using the same study design and assay to assess their utility in treating the disease state, under the assumption that a statistically significant effect in this biomarker would therefore be somewhat predictive of a clinically relevant effect on outcomes in later, larger clinical trials.

In this particular study, patients with the disease condition were randomized to receive regimen D or P (drug or placebo), and the biomarker of interest `r` was measured across time on Day 1 (a baseline day) and just prior to and after dosing with drug or placebo at time 0 hours on Day 2. Data for the first subject (102) may be found in Table 9.9. The remainder of the data set may be found on the website accompanying this book. Note that dosing in this setting was conducted double blind (neither the patients, medics, or personnel conducting the assay knew to which treatment the patients were assigned). This prevents the potential introduction of bias, and such a procedure is recommended for such methodology experiments.

Table 9.9 *Example 9.4.1: Biomarker Data from a Methodology Study for Subject 102*

Subject	Day	Time	Regimen	R
102	1	-1	D	10.1
102	1	-0.5	D	10.1
102	1	0	D	8.2
102	1	0.5	D	30.6
102	1	1	D	22.8
102	1	1.5	D	13.8
102	1	2	D	19.9
102	1	2.5	D	20.3
102	1	3	D	14.4
102	1	3.5	D	12.8
102	1	4	D	18.6
102	1	4.5	D	10.6
102	1	5	D	8.4
102	1	5.5	D	8.4
102	1	6	D	42.2
102	2	-1	D	9.0
102	2	-0.5	D	9.1
102	2	0	D	9.1
102	2	0.5	D	24.3
102	2	1	D	19.6
102	2	1.5	D	15.4
102	2	2	D	22.3
102	2	2.5	D	16.4
102	2	3	D	30.1
102	2	3.5	D	10.1
102	2	4	D	14.4
102	2	4.5	D	13.5
102	2	5	D	6.8
102	2	5.5	D	7.2
102	2	6	D	7.7

D=Dose of Drug on Study Day 2

Our approach to data analysis in this setting is similar to that used in Chapter 8 for testing on QTc. Here, there are repeated measures responses over time with a corresponding baseline assessment in each subject. Our assumption (null hypothesis) is that drug does not affect the biomarker, i.e. that,

$$H_0 : \mu_{Ti} - \mu_{Pi} = 0 \tag{9.1}$$

for all i where i denotes biomarker samples collected over the relevant times of sampling. This hypothesis is to be tested versus the alternative hypothesis:

$$H_1 : \mu_{Ti} - \mu_{Pi} < 0 \tag{9.2}$$

for at least one i in the sampling interval.

As we know from Chapter 8 (see Technical Appendix), the false positive rate of such a test is relatively high (certainly in excess of the 5% chance we desire). Therefore, we will employ the Westfall SimIntervals procedure [449] to constrain the occurrence of false rejection of the null hypothesis to 5%.

The SAS code one may use to model such data is as follows.

```
proc mixed data=method
    method=reml ITDETAILS CL
    scoring=50 maxiter=200;
class subject time regimen;
model r=rb regimen time regimen*time
    /DDFM=KENWARDROGER S outp=out;
repeated time/type=AR(1) subject=subject;
    lsmeans time*regimen/corr cov;
ods output LSMeans=LSmeans; run;
```

As in the earlier examples, here the procedure `mixed` is called in SAS, and told to use REML modelling, to print the iterations (`ITDETAILS`), and to do a maximum of 200 iterations (the `maxiter` statement). The `class` describes the descriptor variables of the data set appropriate to the design, and `time` denotes the time of biomarker sampling relative to dosing at time 0 hours on Day 2. The endpoint `rb` denotes the baseline assessment of `r` on Day 1. The findings are then output to a data set called `LSmeans` which are then utilised in the Westfall SimIntervals procedure [449] to construct comparisons between treatments at each sampling time. If one of the adjusted 90% upper bounds does not include zero, the null hypothesis is rejected, and the biomarker may be useful in future assessments of treatment effect. The results may be found in Table 9.10.

Treatment with drug caused a significant effect on the biomarker `r` only at the 1 hour post-dose time point. Such a finding would indicate

Table 9.10 *Example 9.4.1: Mean Changes (Simultaneous 90% Confidence Intervals) between Regimens Following a Drug Administration in a Methodology Study Corrected for Correlation*

Comparison,Time h	Estimate	Adj. 90% CI
D-P,-1	-1.72	(-9.25,5.80)
D-P,-0.5	-2.77	(-10.30,4.76)
D-P,0	-3.26	(-10.79,4.26)
D-P,0.5	-2.28	(-9.80,5.25)
D-P,1	-13.39	(-20.92,-5.87)
D-P,1.5	-3.29	(-10.97,4.40)
D-P,2	-5.22	(-12.90,2.46)
D-P,2.5	-1.86	(-9.51,5.79)
D-P,3	-3.16	(-10.92,4.59)
D-P,3.5	-2.35	(-10.13,5.43)
D-P,4	-5.23	(-13.01,2.55)
D-P,4.5	-2.38	(-10.16,5.39)
D-P,5	-5.82	(-13.74,2.11)
D-P,5.5	-4.56	(-12.49,3.36)
D-P,6	-4.00	(-11.79,3.79)

D = Drug Treatment
P = Placebo

that while the biomarker has the potential to be used in drug development, the assay may need to be improved to provide more precision before utilizing this approach on other drugs.

We encourage readers interested in further application of these methods to explore this data set `method.sd2` on the website accompanying this book. Complete code for the analysis described above is provided in `method.sas`.

CHAPTER 10

Population Pharmacokinetics

Introduction

I was sitting in my office one day minding my own business (i.e., staring out the window) when I received a call from one of our clinical research scientists. I refer to it as resting one's eyes - staring out the window, that is. After staring at statistical outputs of a computer screen all day, it is good to dwell on distance for just a moment or two - if for no other reason than to keep your eyes from going bad.

If anyone gives you a hard time about it, hand them a stack of statistical outputs needing sorting out, review, and interpretation, and ask them to come back to you in two to three hours if they still really have a problem with it. They will not likely come back, and it is possible you will never see them again.

The scientist had received a message from one of our company's offices in the Far East (South Korea), requesting assistance with a statistical issue. It related to one of our key drug projects and was, to paraphrase, 'How does one go about statistically analysing pharmacokinetic data? We just did a study and do not know what to do with the data.'

I was tempted to tell her I did not know either (and to call someone else), but I knew I could not get away with that.... It was my drug project; I did know how to analyse pharmacokinetic data; and even if I referred her to someone else in the company, eventually the question would make its way back to me. I was the one with the Western pharmacokinetic data to which they would wish (even though they did not know it yet) to compare these new data. I must admit I was tempted, though.

What started off as a seeming annoyance, turned into a very interesting project as we began looking at the data that had been generated in South Korea, and we will discuss the statistical assessment of population pharmacokinetics at some length in this chapter. This information is generally used in the label of new drug products to ensure they are used safely and effectively in different populations. Some aspects also may impact regulatory approval of drugs.

10.1 Population and Pharmacokinetics

Beginning in the latter half of the 20th century (as computational tools became available to support its development), study of extent and rate of exposure began and, has since, become the norm in drug development. This study is targeted toward achieving an understanding of the differences in the way disease-bearing patients' bodies handle drug once a dose is taken. It is hoped that this understanding will aid in the determination and control of safe and effective dosage regimens. Most pharmacokinetic methods applied in pharmaceutical development are non-compartmental (see Chapter 2) in that the concentration of drug in plasma or blood over time is expressed as a summary measure (e.g., AUC or Cmax).

The model-based study of population pharmacokinetics is, however, a relatively recent innovation in drug development and is more of an art than a science at this time. Such techniques apply models to describe the population-specific behavior of concentration in plasma or blood as a function of dose over time. The relationship of concentration to population-specific factors is observational.

Dose is varied among populations, and the resulting pharmacokinetic measurements are quantified using models. Except in selected studies (discussed later in this chapter for the purposes of model validation), control of population specific factors is not all that robust. Such studies are designed for other purposes (e.g., safety evaluation), and pharmacokinetic data are collected in case this can help explain any findings of concern (or benefit). While dose is controlled, and can therefore be considered to affect or cause study outcomes, population and demographic factors are not as robustly controlled and can be termed to be **associated** with or **related to** study outcomes, not a direct cause. The purpose of this chapter is to describe procedures used to study this association between population and pharmacokinetics.

We will not review this topic in great detail and refer interested readers to summaries of this topic in [12], [37], [128], and [286]. Instead we will utilize the pharmacokinetic concentration data from Section 7.3 to review concepts in population pharmacokinetic modelling to enable an understanding of the statistical issues involved in this topic of drug development. We will continue to use the first-order compartmental model introduced in Section 7.3 as its properties lend itself to transparent interpretation. More complex models, however, are likely to improve model fit, and we encourage interested readers to examine `conc.sd2` (found on the website accompanying this book) to do so.

Statistically, the study of population pharmacokinetics may be viewed as a modelling exercise. Pharmacostatistical modelling follows several stages in this setting:

1. Model **building** based on the rich concentration data obtained from limited numbers of subjects in Phase I,
2. Statistical and practical model **assessment**,
3. Model **application** as sparse concentration data are obtained in large numbers of patients in Phases II and III,
4. **Utilization** of model estimates for labelling purposes.

We will briefly review the building and statistical assessment of an example model as illustrated in Section 7.3. Recall the concentration data for Subject 47 presented in Table 7.11 (Section 7.3). These data (and data from the other 26 subjects in `conc.sd2`) were used to develop a pharmacostatistical model to describe the concentration versus time profile (see Figure 7.4 and Table 7.12). Readers will recall that model diagnostics revealed that concentrations appeared to be underestimated at low and high concentrations in this model. We now examine the practical implications of this in more detail.

From the first-order model, it is easy to derive model-based estimates for Tmax, Cmax, and AUC with accompanying confidence intervals and to compare them to the non-compartmental estimates derived using the standard techniques described in Chapter 2. SAS code for doing so in this model may be found below. Details of the derivations may be found in the Technical Appendix to this chapter. For the purposes of this example, we will examine how the model-estimated AUC differs from the non-compartmental derived AUC. Similar procedures may be used to examine Cmax, and we encourage interested readers to use the code found on the website accompanying this book to do so. Intuitively, if the model is accurate, the estimates of AUC and Cmax from the model should approximate those found using non-compartmental methods of derivation.

The SAS code below utilizes the model of Section 7.3 to derive estimates of AUC. The code then outputs these AUC values (with confidence intervals) and compares them to the non-compartmental derived AUCs (see Table 10.1). It was found that the estimated AUCs from the model were approximately 20% lower than those derived using the non-compartmental analysis (based on the findings for the ratio of non-compartmental AUC to model-based AUC).

Derivation of Tmax, Cmax, and AUC from Nonlinear Mixed Effect Pharmacokinetic Data Analysis of `conc.sd2` *- SAS* `proc nlmixed` *code:*

```
proc sort data=my.conc;
    by subject dose time;run;

proc nlmixed data=my.conc;
    parms beta1=0.4 beta2=1.5 beta3=-2 s2b1=0.04
    s2b2=0.02 s2b3=0.01 s2=0.25;
    cl = exp(beta1+b1);
    ka = exp(beta2+b2);
    ke = exp(beta3+b3);
    auc=dose/cl;
    tmax=(log(ka)-log(ke))/(ka-ke);
    pred=dose*ke*ka*(exp(-ke*time)-exp(-ka*time))/
    (cl*(ka-ke));
    cmax=dose*ke*ka*(exp(-ke*tmax)-exp(-ka*tmax))/
    (cl*(ka-ke));
    model conc ~ normal(pred,s2);
    random b1 b2 b3 ~ normal([0,0,0],[s2b1,0,
    s2b2,0,0,s2b3]) subject=subject;
    predict auc out=auc;
    predict cmax out=cmax;
    predict tmax out=tmax;
    run;

data auc;set auc;if time=1;run;
    proc sort;by subject;run;
data auc_nc(keep=subject auc_nc); set my.pk;
    if day='single';auc_nc=auc; run;
    proc sort;by subject;run;

data auc;merge auc auc_nc;by subject;
    diff=pred-auc_nc;
    ratio=auc_nc/pred; run;

title 'Difference in AUC Comp versus NC';
  proc print data=auc;
  var subject dose pred lower auc_nc upper diff ratio;
  run;
  proc univariate data=auc;var diff ratio;run;
```

The explanation for this discrepancy in estimates, in this manufac-

Table 10.1 *Estimated AUC Parameters from* `conc.sd2`

Subject	Dose	Model AUC	Model Low B.	Non-Comp AUC	Model Upper B.	Diff.	Ratio
47	5	3.24	2.29	2.81	4.18	0.43	0.87
48	5	4.36	3.24	6.31	5.48	-1.95	1.45
49	5	4.80	3.59	7.26	6.00	-2.46	1.51
50	5	3.51	2.54	3.60	4.48	-0.09	1.03
52	5	4.23	3.10	6.82	5.37	-2.59	1.61
53	5	2.85	1.96	1.76	3.75	1.09	0.62
54	5	4.83	3.62	6.11	6.05	-1.28	1.26
55	5	3.93	2.87	6.09	5.00	-2.16	1.55
57	5	3.24	2.30	2.10	4.18	1.14	0.65
60	10	7.63	6.16	9.33	9.11	-1.70	1.22
61	10	6.45	5.12	7.31	7.78	-0.86	1.13
62	10	7.16	5.71	9.57	8.60	-2.41	1.34
64	10	8.45	6.83	15.62	10.07	-7.17	1.85
65	10	5.58	4.25	5.56	6.91	0.02	1.00
66	10	6.34	4.90	11.81	7.78	-5.47	1.86
69	10	7.36	5.93	7.23	8.80	0.13	0.98
71	10	6.68	5.30	8.35	8.07	-1.67	1.25
72	10	6.02	4.74	5.70	7.31	0.32	0.95
95	20	13.65	11.44	12.92	15.86	0.73	0.95
99	20	19.56	16.45	26.05	22.67	-6.49	1.33
102	20	18.32	15.60	23.12	21.05	-4.80	1.26
104	20	11.91	9.94	12.32	13.87	-0.41	1.03
105	20	13.16	11.05	16.35	15.27	-3.19	1.24
106	20	15.43	13.03	20.21	17.83	-4.78	1.31
107	20	11.12	9.19	13.53	13.05	-2.41	1.22
108	20	9.53	7.64	7.70	11.42	1.83	0.81
110	20	12.22	10.19	14.22	14.25	-2.00	1.16

tured example, is as follows. Interested readers will recall that in theory (see Section 7.3)

$$AUC = \frac{F(Dose)}{Cl},$$

where F denotes the ratio of absolute bioavailability (see Section 10.3). No basis for the derivation of this F is present in this data set (as no intravenous route of administration was included in the study). In science, such ‘fudge-factors’ are often employed while learning about the science to account for differences in model estimates to actual observations (e.g., Einstein’s cosmological constant [54]), and we will utilize this procedure here for the purposes of illustration. In practice, input from a pharmacokineticist should be sought to determine what procedure for adjustment should be used or if another model should be built and assessed. For the purpose of illustration, we adjust the model estimated AUC by a factor of 1.2 using the following SAS code accordingly:

```
auc=1.2*dose/cl;.
```

Based on the model parameters, and our rough estimate for F, we now have a model-based means of constructing accurate AUC estimates from concentration data (for illustration purposes). Subsequent Phase I studies collect more concentration data to enhance the understanding of the model, and at the end of Phase I, a more robust model should have been developed relating clearance (etc.) and dose to AUC and Cmax.

It should be expected that the building of a model and statistical and practical assessment of its properties is an iterative and collegial process. Such models are built by statisticians and pharmacokineticists in consultation with disease area experts and their medical colleagues. Those building and assessing such models should bear in mind George Box's statement 'All models are wrong, but some are useful.' [40] The idea is to build and assess a parsimonious model describing the data adequately. Adequacy of model fit and performance is to some extent subjective.

Turning now from these topics, we now consider the application of a model to emerging clinical pharmacokinetic data obtained in Phase II and III patient studies. Such data are generally more sparse than Phase I data (in that a full pharmacokinetic profile sufficient for estimation of AUC and Cmax is not obtained); however, these sparse collections are generally obtained in a far larger number of patients than were exposed to drug in Phase I. Selected data for three subjects may be found in Table 10.2. The full simulated data set may be found in `simulate.sd2` on the website accompanying this book.

Table 10.2 *Selected Sparse Concentration Data from Patient Studies*

Subject	Dose	Time	Concentration
1	5	1	0.21
1	5	3	0.18
1	5	6	0.13
1	5	14	0.05
40	10	1	1.10
40	10	3	0.95
40	10	6	0.74
40	10	14	0.38
90	20	2	0.79
90	20	5	0.62
90	20	8	0.48
90	20	18	0.20

These data are concentrations from three of 100 simulated patients. Note that the number of concentrations obtained are limited relative to the normal healthy volunteer data (Table 7.11). Using the model developed in Phase I, we use the above SAS code to derive parameter and AUC estimates for each subject. The same code as above is used except that the starting values are based on the findings from the Phase

I model in Table 7.12 using a PARMS statement of: `parms beta1=0.35 beta2=1.46 beta3=-2.47 s2b1=0.04 s2b2=0.03 s2b3=0.007 s2=0.01;`

Code for this purpose may be found in `poppk.sas` on the website accompanying this book. Model diagnostics may be applied (although not done for the purposes of this example), and if model fit is poor, alternative models may be built and assessed. Parameter estimates may be found in Table 10.3 (note slight differences from the Phase I estimated parameters in Table 7.12), and resulting AUC estimates for selected patients may be found in Table 10.4.

Table 10.3 *Estimated Population PK Parameters from Sparse Population Data*

Parameter	Estimate	95% CI
β_1	0.45	0.39,0.52
β_2	1.47	1.11,1.83
β_3	-2.44	-2.48,-2.39
s2b1	0.10	0.08,0.11
s2b2	0	.,.
s2b3	0.03	0.02,0.04
s2	0.0003	0.0002,0.0004

As shown in Table 10.4, the estimates of AUC (and the other parameters) have uncertainty (error) associated with their estimation. In SAS, a Bayesian algorithm [368] is applied to characterize this uncertainty. In theory, the bootstrap may also be applied (in addition to its use as a model diagnostic to assess model performance) to provide an estimate for the uncertainty of the estimate; however, this is generally not done given constraints on modern computing power. Nonlinear mixed effects models of the type described often take several minutes or hours to run, and if 1000 bootstrap runs (see Chapter 5) are performed, very lengthy computation time can result.

We turn now to the utilization of these estimates from the model. Estimates of AUC and clearance for our 100 simulated patients may be found in `pop_auc.sd2` and `pop_cl.sd2` on the website accompanying this book. The first goal is to use the estimated AUC to confirm their position relative to the NOAEL in this population. The estimated AUCs are plotted versus dose in Figure 10.1.

Similar procedures may be done for the estimated Cmax, and we leave this as an exercise for interested readers.

Table 10.4 *Selected Estimates for AUC from Sparse Concentration Data Obtained in Patient Studies*

Subject	Dose	AUC	95% CI
1	5	3.00	2.37,3.63
40	10	17.10	16.03,18.17
90	20	13.09	12.17,14.01

The second goal of population pharmacokinetic analysis is to assess the estimated parameters (in this case we will use clearance) relative to factors which may influence their magnitude. Examples include dose and demographic factors such as age, gender, weight, body-mass-index, ethnicity, and creatinine clearance (a measure of renal function). Basic statistical tools are often used to enable assessment of whether changes in these factors influence the magnitude of the estimated population pharmacokinetic parameters, see Figure 10.2.

Figure 10.2 is a plot of the estimated clearance (from the model) versus dose expressed using a standard descriptive statistical procedure known as a *box-plot*. The box encloses the 75th and 25th quartiles of the observed data, and the line in the box is the median of the observed data. The upper and lower lines extend to the 90th and 10th quartiles, respectively, with data outside these indicated using points so their status as outliers can be assessed.

In Figure 10.2, we observe that clearance appears related to dose. This relationship may be further quantified by regressing the estimated clearance on dose to assess whether the relationship is linear or nonlinear. Multiple linear regression may be performed to assess the simultaneous relationship of other (i.e., demographic) factors [314]. We will not dwell further on such assessments here and refer interested readers to discussion in Chapter 11 of [37] for more details.

Such model-based population pharmacokinetic assessments are used to guide dosing in patients where well-controlled clinical designs are not possible (e.g., [138]) due to ethical or practical constraints. Additionally, this information will be used in labelling for the drug product [128] to ensure dosing of patients in the marketplace is appropriate to their demography and concurrent-disease states.

Exposure levels above the NOAEL or exposure levels related to a demographic factor which may be impacted by a concurrent-disease state may be the subject of specific clinical pharmacology studies to assess the relationship of exposure to disease or demography. Following a brief dis-

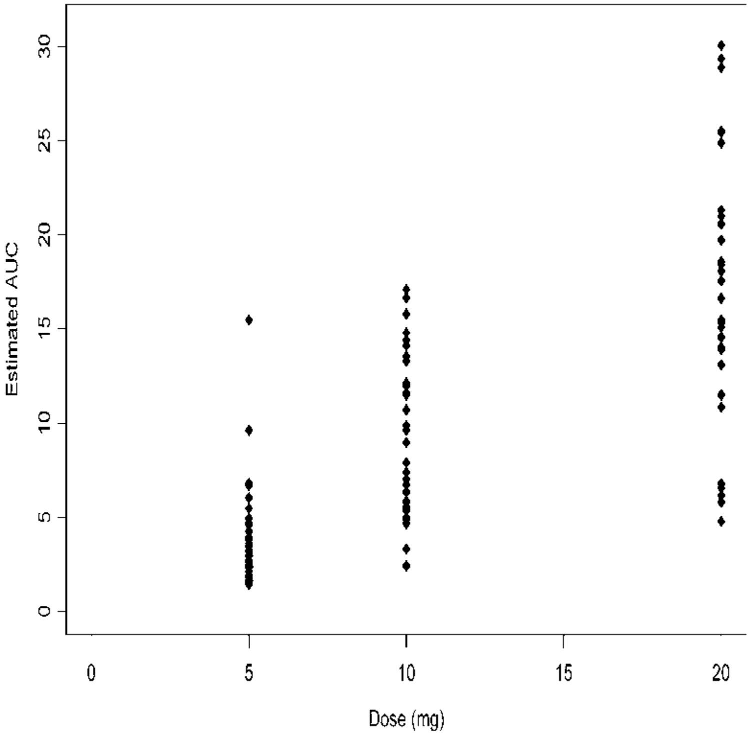

Figure 10.1 *Estimated AUCs versus Dose from a Simulated Population Pharmacokinetic Study*

cussion of the determination and estimation of absolute bioavailability, we turn to several examples of such studies.

10.2 Absolute and Relative Bioavailability

As described in Chapter 1, when a drug is taken orally, it is absorbed and distributed into the body, metabolized at various sites within the body, and eventually eliminated from systemic circulation. This process is termed 'ADME,' and the availability of drug at the site of action within the body is presumably mediated by the rates at which the various facets of ADME are performed by the body.

Consider, however, a drug that is injected or administered intravenously. Once administered, the drug is distributed to the systemic circulation from the site of entry and does not undergo first-pass metabolism as do

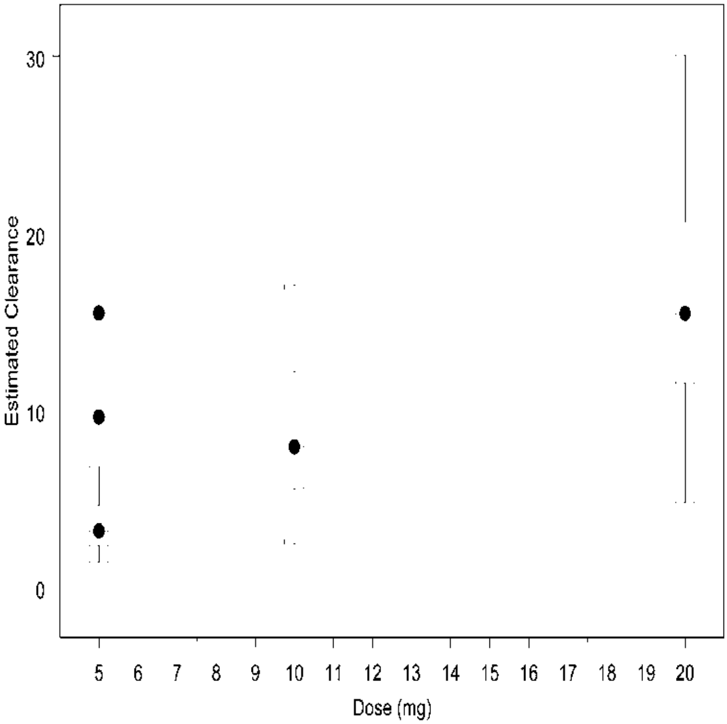

Figure 10.2 *Box-plot of Estimated Clearance versus Dose from a Simulated Population Pharmacokinetic Study*

drugs which are ingested when they are absorbed through the intestinal tract through the liver. As the injected drug product circulates throughout the body, it is metabolized and eliminated. Equation (10.1) is appropriate for such a product (see Technical Appendix his is termed 100% bioavailable as an injected product by definition reaches the circulation intact at the time of dosing. Most oral products have different levels of bioavailability as some drugs pass straight through the intestinal tract and are eliminated, and some drugs (like the example of the previous section) can be very rapidly absorbed in the intestinal tract. To account for this in equations like (10.1), parameters such as F can be introduced to account for the differential mode of administration (Chapter 7, [37]).

Description of absorption pharmacokinetics is a lengthy topic, and we will not discuss all aspects of its assessment. Instead, we will discuss a commonly used method to assess absolute bioavailability F using data from a cross-over clinical pharmacology trial. Such a trial need not always be performed in drug development. In certain circumstances, F can be determined by other means (see Chapter 7, [37]).

Absolute bioavailability F is a measure of the percentage of drug absorbed after oral administration relative to that in the body after administration by an intravenous route (hereafter denoted IV). This parameter F can be estimated by giving an IV dose and an oral dose of drug in a cross-over study to normal healthy volunteers and comparing their resulting AUCs.

The same approach to study design is used as in the typical bioequivalence study; however, here we do not desire to demonstrate equivalence in AUC but only to estimate F to a given degree of precision. Usually, the dose of drug administered IV and orally in such trials will differ depending on the properties of the compound to ensure that exposure levels remain safe. For example, a drug poorly absorbed after oral adminstration might have a reduced dose when administered IV to ensure concentrations remain below the NOAEL. Therefore, the AUCs are dose-normalized (i.e., divided by dose) prior to analysis to ensure that an appropriate basis for comparison is obtained. A SAS program to determine precision in F, for a given sample size, in 2×2 cross-over designs is provided in the Technical Appendix.

Table 10.5 contains data from a typical cross-over trial to estimate absolute bioavailability. In this case, 2 mg of drug was administered intravenously over an hour or 4 mg of drug was administered orally in a cross-over trial in $n = 12$ normal healthy subjects, and dose-normalized AUC values were derived following each administration.

The dose-normalized data of Table 10.5 were analyzed according to the methods of Chapter 3 (SAS code may be found on the website accompanying this book), and an estimate of $\mu_O - \mu_{IV}$ with a 90% confidence interval was constructed (where μ_O and μ_{IV} denote the adjusted mean logAUC following oral and IV administration, respectively). As with bioequivalence, these are exponentiated to provide an estimate of F. In this case $\hat{F}$ was 0.99 with 90% confidence bounds of 0.91 to 1.07.

Information provided by the models of this and the previous section and Chapter 7 are necessary but not always sufficient for complete understanding of the ADME properties of a drug. To complete the scientific understanding of ADME properties, a single dose, cross-over mass-balance (see Chapter 5, [37]) study is often performed in an extremely small number of normal healthy volunteers ($n = 2$ to 4 total). In such trials, subjects are administered a radio-labelled dose of drug, and blood and other bodily excretions (urine, feces) are collected and assessed for the presence of a radio-labelled substance. In the other session, a standard dose of drug is given to serve as a control for the amount of drug

Table 10.5 *Dose-Normalized (DN) AUC from an Absolute Bioavailability Cross-over Trial*

Subject	DN-AUC *IV*	DN-AUC Oral
1	751	818
2	897	694
3	900	954
4	537	469
5	656	665
6	665	681
7	772	578
8	930	869
9	884	1055
10	556	506
11	1029	1078
12	727	946

(and radio-label) found in blood in the other session. Pharmacokinetic data from such a trial are generally not analyzed statistically (given the low sample size) but are used qualitatively to confirm the scientific understanding of the ADME properties of drug products. As such, we do not consider their statistical properties here.

During the early stages of drug development, many changes are made to formulation. These may be minor (changing the color) but can be major (e.g., changing from a capsule to a tablet). Guidance [135] does not require that a bioequivalence trial be performed, but sponsoring companies will wish to confirm that AUC is similar in the new formulation to ensure that the understanding of absolute bioavailability gained in previous experimentation is robust to the change in formulation.

As with absolute bioavailability studies, bioequivalence need not be demonstrated, and such relative bioavailability trials are performed to provide the desired level of precision in the ratio of AUC in the new formulation to the old. Study design and data analysis follow the same principles of those used in bioequivalence testing as described in Chapter 3 and will not be discussed further here. SAS code to support the assessment of the desired level of precision for given sample sizes may be found in the Technical Appendix.

10.3 Age and Gender Pharmacokinetic Studies

As described in Section 10.1, population pharmacokinetic models will be used to relate clearance and other pharmacokinetic parameters relative to age and gender. Such models, however, are handicapped with decreasing confidence as findings are extrapolated beyond the observed data [314].

For example, clinical trials of a new drug product may only be done in adults (ages 18 years to 50 years, perhaps). The models of Section 10.1 will allow for extrapolation to lower and higher ages (down to zero and up to, say, 100+ years perhaps); however, the confidence in the model predictions decreases as distance from the observed age range increases. Of interest, then, would be how exposure will actually behave in very young people or perhaps very old people? Age pharmacokinetic studies are designed to go and check. As noted, estimates will be available from the population pharmacokinetic model, and often a limited pharmacokinetic study is performed to assess whether these model estimates are dependable. These small age (and gender) studies are, in essence, model-validation tools.

Consider the data in Table 10.6 from a study where pediatric patient pharmacokinetics were assessed for such a purpose. Ten pediatric and ten adult subjects received a single dose of drug, their plasma concentrations were measured in the usual fashion over time.

Note that age, weight, and height were expected to differ between the two age groups, but these were not related to clearance and concentration in the population pharmacokinetic models (data not shown). Weight, height, kidney, and liver function all differ also (hopefully for the better in the younger people).

The study was performed to assess whether exposure in juveniles was consistent with this finding. The resulting findings are observational. Demographic characteristics will differ between groups, and the adult subjects are included, not for purposes of direct comparison, but to serve as a control back to the model used in the population pharmacokinetic modelling population. Their inclusion serves as a control if unexpected findings are observed to determine if the model or some facet of the study (e.g., assay) explains the observed difference.

The objective of statistical analysis in such a setting is to estimate exposure levels with desired level of precision and compare to the NOAEL, calibrating back to estimates from the population pharmacokinetic models. SAS code commonly used to do so follows.

Table 10.6 *AUC and Cmax Data from a Pediatric (PED) and Adult (ADT) Bioavailability Trial*

Subject	Age	AUC	Cmax
201	PED	1510	88.6
202	PED	883	52.5
203	PED	1650	92.0
204	PED	1015	56.0
205	PED	1556	84.0
206	PED	1412	84.8
207	PED	1353	83.0
208	PED	1443	96.4
209	PED	1299	68.1
210	PED	560	33.5
101	ADT	1284	70.3
102	ADT	1391	73.5
103	ADT	873	50.2
104	ADT	1211	62.2
105	ADT	1233	74.1
106	ADT	1172	60.4
108	ADT	1172	60.4
109	ADT	1336	75.3
110	ADT	1348	76.8
112	ADT	1419	82.9

Age AUC Assessment Example - SAS `proc mixed` *code:*

```
proc mixed method=reml data=age;
    class subject age;
    model lnauc=age/
    s ddfm=kenwardroger cl alpha=.1;
    lsmeans age/cl alpha=0.1;
    repeated /group=age subject=subject;
    ods output LSMeans=auc;
    run;
```

As previously, a REML model is used to characterize the mean AUC and Cmax of such data. As this is a parallel group trial, the model simply calls for characterization of logAUC relative to age, with mean effect in each age group output in the `lsmeans` and `ods` statements. The `repeated` statement specifies that a variance estimate should be provided for each age group separately as it may be unrealistic to expect variation to be the same in the pediatric population relative to the adult population.

Exponentiating the estimated means and 90% confidence intervals back to the natural-scale, it was found that mean AUC in the adult subjects was 1234 (1138-1338) and in the pediatric subjects was 1214 (1000-1474). These estimates were as expected from the population pharmacokinetic modelling and served to reassure those using the drug that the choice of dose in this population was safe relative to the NOAEL.

Similar to pediatric subjects, for elderly people, it will often be of interest to assess the findings of population pharmacokinetic models in this manner. As before, weight, height, kidney, and liver function all differ too (probably for the worse in the older people). We omit further discussion on this topic here as the principles and analyses are similar to those used in the pediatric population.

A particularly important facet of the application of population pharmacokinetic data pertains to assessment of the relationship of gender to exposure levels. As discussed in Chapter 8, females are more likely to experience undesired side effects than males. In population pharmacokinetic models, interpretation of gender's relationship to exposure is often not straightforward. Confounding with other demographic factors is significant - i.e., weight and height. In general, to obtain a good handle on whether exposure is gender related, a single-dose study of exposure levels relative to NOAEL is done early in drug development (usually just after the sub-chronic dosing study, see Section 7.3, completes) in Phase I. Inclusion of females of child-bearing potential in drug development studies is contingent on genotoxicology findings and protocol contraceptive procedures as fetal development can be impaired or terminated by such products. Effort is made to weight match male and female volunteers from the different populations where possible.

Consider the AUC and Cmax data from a gender trial in Table 10.7. Here 18 males and females (six per dose) were given a single dose of drug, and their pharmacokinetics were measured in the usual fashion.

Table 10.7 *AUC and Cmax Data from a Gender Bioavailability Trial*

Subject	Dose	Gender	AUC	Cmax
1	1	M	354	68.4
2	1	M	219	50.5
3	1	M	228	36.5
4	1	M	216	55.6
5	1	M	405	74.6
6	1	M	306	55.5
13	1	F	704	90.0
14	1	F	375	52.3
15	1	F	534	83.7
16	1	F	434	59.2
17	1	F	565	59.8
18	1	F	484	84.0
25	2	M	602	151.5
26	2	M	762	165.6
27	2	M	728	134.6
28	2	M	934	116.6
29	2	M	560	121.2
30	2	M	408	86.9
38	2	F	871	196.3
39	2	F	1104	216.0
40	2	F	777	80.1
41	2	F	592	109.7
42	2	F	728	122.5
49	5	M	2295	553.5
50	5	M	1743	307.8
51	5	M	1646	483.4
52	5	M	1523	281.4
53	5	M	1782	534.4
54	5	M	1906	375.0
61	5	F	1676	211.8
62	5	F	1493	266.7
63	5	F	2597	328.1
64	5	F	2396	242.3
65	5	F	1656	455.4
66	5	F	1355	288.8

As with the pediatric trial described above, statistical analysis of the pharmacokinetic data is geared toward providing estimates which may be used to calibrate the population pharmacokinetic findings. SAS code for this purpose follows.

Gender AUC Assessment Example - SAS `proc mixed` *code:*

```
proc mixed method=reml data=gender;
    class subject dose gender;
    model lnauc=dose gender dose*gender/
    s ddfm=kenwardroger cl alpha=.1;
    lsmeans dose*gender/cl alpha=0.1;
    repeated /group=gender subject=subject;
    ods output LSMeans=auc;
    run;
```

Again, variability is allowed to differ between genders using a `repeated` statement, and mean logAUC is output from the `lsmeans` and `ods` statements. Exponentiating these findings back to the normal-scale, the estimates given in Table 10.8 were found.

Table 10.8 *Mean AUC Findings from a Gender Bioavailability Trial*

Dose	Gender	Mean AUC	90% CI
1	Female	506	426-600
2	Female	797	661-962
5	Female	1808	1523-2146
1	Male	279	235-332
2	Male	644	541-766
5	Male	1800	1514-2142

Relative to the population pharmacokinetic model findings (based up to this time on data from male subjects only), we see in Table 10.8 that while mean AUC in females still falls below the NOAEL (greater than 2000 ng.h/mL for this drug at this time), average exposure in females was dramatically greater in this data set at lower doses than would be expected from the models of male data.

Findings such as these would prompt the sponsor to reinterrogate the population pharmacokinetic model building and assessment procedures (of Sections 7.3 and 10.1), and the concentration data of the gender trial would be utilised for this purpose. Using such techniques, it was determined that, unexpectedly, clearance was related to weight (data not shown). This enabled the team to adapt their population pharmacokinetic model to take this into account. For example, in the SAS code of Section 10.1, `beta1` might be defined as a function of weight where relevant parameters are determined from model-based regression of weight on estimated clearance.

Assessment of Cmax in age and gender trials are left as an exercise for interested readers, and SAS code to perform such analysis may be found on the website accompanying this book. SAS code to aid in the determination of sample size for age and gender pharmacokinetic trials may be found in the Technical Appendix.

10.4 Ethnicity

Consideration of ethnicity's impact upon pharmacokinetics has long been a topic of discussion and was recently commented on in international regulatory guidance [227]. This ICH-E5 guidance [227] was intended to provide a framework for evaluating ethnic factors on a drug's efficacy and safety profile in drug development. At this time, however, the guidance has yet to be fully implemented in the local ICH regions (USA, Europe, and Japan), and there is still a great deal of question about how to interpret the guidance (e.g., [139], [315]).

ICH-E5 [227] makes the implicit assumption that registration of a drug in a new region involves new registration for a new ethnic population, and we will follow this convention in this section. As described in ICH-E5, the first of two primary requirements for a submission package is that the data requirements for registration in the new region be met - i.e., that clinical trial methodology, recordkeeping, protocol compliance and drug accountability, and informed patient consent must be acceptable in the new region [227]. The minimal data package, consisting of either data from the original region and/or data from the new region, should include an adequate characterization of the pharmacokinetics (PK), pharmacodynamics (PD), dose response, efficacy and safety of the drug (see Chapter 1 for more details). At least pharmacokinetics [316], and preferably pharmacodynamics and dose response, should also be characterized in an ethnic population that is relevant to the new region [227] but not necessarily resident in the new region [317] (i.e., if one wants to market a drug in Japan, one has to study its properties in Japanese patients or in patients of Japanese descent).

The second requirement is the demonstration of the ability to extrapolate findings from any data from the original region to the population of the new region. It is easier to extrapolate from one region to another if the new medication is 'ethnically insensitive,' i.e., unlikely to behave differently in different populations. Ethnic sensitivity can be categorized into two components, intrinsic (genetic) and extrinsic (environmental), either or both of which may impact bioavailability and hence the appropriate dose and response relationship. These are described in greater detail in Figure 10.3.

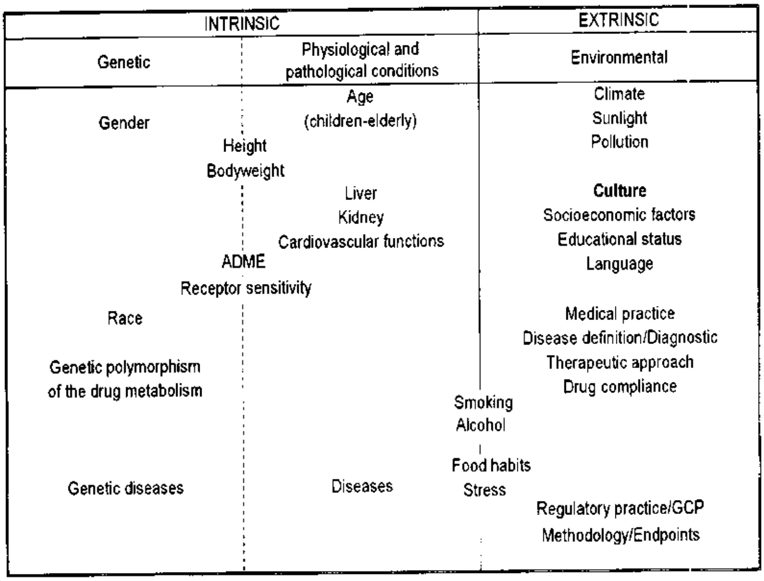

Figure 10.3 *Intrinsic and Extrinsic Population Factors Impacting the Dose-PK-PD Response Relationship [227]*

A ‘bridging’ study, as its name implies, is designed to allow one to bridge from the original region’s data in the original population to the new region with its new population. It is a [227]:

> ....supplemental study performed in the new region to provide pharmacodynamic or clinical data on efficacy, safety, dosage, and dose regimen in the new region that will allow extrapolation of the foreign clinical data to the new region...

The degree of ethnic sensitivity will determine whether a study is necessary and the design of such a study (e.g. PK only, PK/PD only, in what population, etc.). ICH-E5 [227] describes several characteristics of drug products which would make such a product ‘ethnically insensitive.’ These are [227]:

1. Linear pharmacokinetics
2. A flat response curve for both efficacy and safety in the range of the recommended dosage and dose regimen (this may mean the medicine is well tolerated)

3. A wide therapeutic dose range (again an indicator of good tolerability)
4. Minimal metabolism or metabolism distributed among multiple pathways
5. High bioavailability, thus less susceptibility to dietary absorption effects
6. Low potential for protein binding
7. Little potential for drug-drug, drug-diet, and drug-disease interactions
8. Nonsystemic mode of action
9. Little potential for inappropriate use

It is rare for a drug to meet all nine conditions which would make it ethnically insensitive and result in only minimal data requirements to enter new regions and markets (e.g., such as Asia). In any event, ethical and cultural considerations regarding drug use in Asia are slightly different than other international regions, and consideration should first be given to such matters (regardless of the outcome of this checklist) when designing a bridging program [424].

Statistical approaches to bridging are in early stages of development, and no international consensus is yet available on how ethnicity bridging programs should be designed and data analyzed. See [244], [404], [367], [278], [77], [279], [334], and [280] for a description of some methods which are publicly available. We will not discuss these approaches further here as they are, in general, intended for application to bridging study data to confirm these are sufficient to permit market access. We turn to practical pharmacology-based ethnicity assessment in population pharmacokinetics and the statistics involved. These pharmacology assessments are usually carried out in drug development prior to the initiation of a bridging program and should constitute the major basis for the approach to its design.

We assume as in previous sections of this chapter that a population pharmacokinetic model has been developed (as in Section 10.1) describing concentration as a function of time and physiologic parameters (e.g., clearance, absorption constant(s), elimination constant(s), etc.) As described in Section 10.1, some of these physiologic parameters may be related to demography (e.g., weight, height, gender, etc).

When dosing a new population, it is to be expected that demographic factors may be different. As with the population pharmacokinetic assessment of gender, significant confounding with ethnicity can often be expected. For example, in the data set which follows, Western subjects were on average heavier than South Korean subjects. The working assumption, in the absence of information, made in the early stages of model development is that the functional form of the model is the same for both populations; however, in reality the magnitude of parameters (e.g., clearance) may be dependent upon ethnicity in some, as yet unknown, way.

Once a population pharmacokinetic model is proposed and estimates are available for differences in pharmacokinetics between populations related to demographic factors, the logical next step is to conduct a validation exercise via a small focused pharmacokinetic study. Selected data from such a study in South Korean subjects are presented below along with corresponding data (at the same doses) observed in Western subjects. The full data set may be found on the website accompanying this book.

Table 10.9: AUC and Cmax Data from a Population Pharmacokinetic Assessment of South Korean and Western Subjects

Dose (mg)	Ethnicity	Subject	AUC (ng.h/mL)	Cmax (ng/mL)	Weight (kg)	Height (cm)	Age (yrs.)
2	K	A01	1228	195.2	65.0	170	20
2	K	A02	1003	193.9	65.0	172	20
2	K	A03	1063	165.8	75.0	175	27
2	K	A04	906	215.2	64.0	172	22
2	K	A07	811	215.6	76.0	177	21
2	K	A08	928	167.5	82.0	187	21
2	K	A09	1401	136.4	65.0	178	29
2	K	A11	1099	206.1	59.0	168	26
2	W	1	746	208.4	73.7	177	28
2	W	1	734	137.7	71.7	175	28
2	W	11	994	190.9	58.7	180	20
2	W	12	552	125.7	87.6	179	38
2	W	13	675	168.7	59.5	163	36
2	W	13	566	104.9	63.6	175	26
2	W	14	637	108.0	91.7	180	28
2	W	15	666	169.4	70.7	162	30
2	W	15	728	167.2	78.0	176	32
2	W	16	578	123.8	76.2	173	36
			....				
4	K	B01	1763	345.6	64.0	174	21
4	K	B02	1638	302.4	68.0	178	25
4	K	B03	1894	345.8	66.0	171	25
4	K	B06	2125	373.2	69.0	182	26
4	K	B07	2289	466.4	63.0	170	26
4	K	B08	1380	336.9	68.0	181	25
4	K	B10	1557	257.2	80.0	180	23
4	K	B11	3035	335.2	57.0	169	24
4	W	1	1637	362.0	76.3	180	27
4	W	10	2109	371.0	71.3	181	30
4	W	104	1468	308.0	76.1	185	51
4	W	109	999	249.0	85.0	178	56
4	W	11	1012	174.0	109.0	200	25
4	W	115	1273	275.0	69.4	175	31
4	W	116	1322	302.0	74.6	168	22
4	W	2	1388	319.0	68.3	175	26
4	W	391	989	174.1	91.1	176	27
			....				
8	K	C01	4890	709.2	64.0	176	19
8	K	C02	3641	737.7	65.0	167	28
8	K	C04	7211	981.7	63.0	175	22
8	K	C06	3382	421.4	68.0	182	21
8	K	C07	5459	1009.0	59.0	171	28
8	K	C08	3077	769.6	71.0	179	28
8	K	C11	4144	820.0	73.0	177	24
8	K	C12	4263	673.0	61.0	180	21
8	W	1	3404	687.1	73.7	177	28
8	W	1	2942	563.6	80.5	184	26
8	W	10	3596	550.4	79.5	173	26
8	W	10	2148	462.0	76.3	182	32
8	W	100	2572	718.0	73.2	175	54
8	W	106	1997	428.0	96.8	186	43
8	W	11	4677	586.1	58.7	180	20
8	W	11	1278	320.0	96.1	178	29
8	W	112	3023	467.0	80.9	173	49
8	W	113	2959	575.0	69.8	170	53
			....				

K= South Korean; W = Western

The code used to analyse such data is similar to that used in the previous section. In this setting, it may be desirable to conduct a model-building assessment (see Chapter 2 of [314] and Chapters 2 and 4 of

[179]) to determine which factors are significantly related to the endpoint under study. Accordingly, in the example that follows, weight (`wt`) was included as a covariate.

Ethnicity AUC Assessment Example - SAS `proc mixed` *code:*

```
proc mixed data=pk method=reml maxiter=200 scoring=50;
    class subject race;
    model lnauc=race lndose wt
    /s ddfm=kenwardroger cl alpha=.1 outp=out;
    lsmeans race/CL ALPHA=0.1 DIFF=CONTROL("W");
    repeated /group=race subject=subject;
    ods output LSMeans=auc;
run;
```

For logAUC, the resulting model estimates are presented in Table 10.10. AUC was observed to be significantly related to ethnicity and weight, and was linearly related to dose. In terms of the impact of ethnicity, we can conclude from these data that weight, by itself, does not explain all of the differences in pharmacokinetics between Koreans and Westerners. The concentration data supporting this assessment would be used to rebuild the population pharmacokinetic model allowing for other parameters to be related to ethnicity.

Table 10.10 *Estimated Population Parameters from Evaluation of logAUC as a Function of Ethnicity, logDose, and Weight*

Parameter	Estimate	95% CI
Ethnicity	0.28	0.19, 0.36
logDose	1.00	0.94, 1.05
Weight	-0.01	-0.02, -0.00

In this case, it was determined that South Koreans metabolized the drug slightly differently than Westerners (via a different CYP450 pathway, see Chapter 7). Alteration of the elimination rate constant to account for this ethnicity-related difference resulted in adequate model fit (data not shown).

Cmax was also observed to be higher in South Koreans than the Western population. The analysis of these data is left as an exercise for interested readers and may be done using code on the website accompanying this book.

In combination with the full data package from the original region, data such as the above can serve as the basis for approval in some nations. However, several nations also require that the concentration to effect relationship (see Chapter 9) be studied and be shown to be not related to ethnicity. In theory, the model-based approach used should be similar, but this is a still evolving topic, and we will not discuss it further here.

Code to determine precision in pharmacokinetic findings for given sample sizes in ethnicity studies is the same as that used in age and gender studies and is omitted here.

10.5 Liver Disease

Liver disease or hepatic impairment can be caused by a number of factors. Diseases like hepatitis can cause injury to the liver and impair its function. Injury may also be chemically induced (cirrhosis via alcohol) and drug induced. In this section, we consider the pharmacokinetics of a drug in the body when patients have liver disease.

Severity of liver disease is typically measured by the Child-Pugh score [136], and subsequently categorized as healthy, mild, moderate, or severe liver function impairment, depending on extent of damage to the liver and impairment of its function. If a drug is eliminated (in the ADME sense) by metabolism or excretion (into bile) in the liver, drug would be expected to accumulate in the plasma. Decreased clearance of drug by the liver [12] implies increased AUC and Cmax, and as these increase the likelihood of adverse events associated with exposure (relative to the NOAEL) would also be expected to increase. Therefore, it is important to understand the magnitude of increased exposure in patients with impaired hepatic function to determine [12] if it is necessary to reduce dose in such patients or potentially to contraindicate the use of the drug.

We again assume that a population pharmacokinetic model has been developed from Phase I data. In tandem with this, the mass-balance radio-label ADME trial (see Section 10.2) will generate information on the role of the liver in excretion and metabolism of the drug in plasma. If the liver plays only a minor role in elimination of the drug from the body [136], [12], then regulatory guidance suggests that study in patients with hepatic impairment is not required. However, if the role of the liver cannot be precisely determined, then a small pharmacokinetic study is generally performed to confirm the validity of the model's findings. In practice, the radio-label ADME study is expensive and takes a long time, so it is general practice to perform a small pharmacokinetic trial as described in the following.

Patients with hepatic impairment are enrolled and administered a single dose of drug in the standard clinical pharmacology sampling paradigm, and their plasma concentrations are summarized as AUC, Cmax, etc.

[136]. In tandem, depending on the results of population pharmacokinetic assessment for the demographic factors involved, race, age, and weight range-matched volunteers are enrolled as a control group, administered the same single dose, and pharmacokinetics are measured. As with previous population pharmacokinetic modelling exercises, the objective of the trial is to estimate the pharmacokinetics in each group to assess the performance of the population pharmacokinetic model, not to compare the groups ('normal' and 'hepatic impaired'). Code to determine precision in pharmacokinetic findings for given sample sizes in hepatic impairment pharmacokinetic studies is the same as that used in age and gender studies and is omitted here.

AUC and Cmax data from such a trial may be found in the following table. In this case, population pharmacokinetic modelling of the impact of reduced clearance due to hepatic impairment led the team working on this drug to be confident that increased extent of exposure would occur in hepatic impaired patients. The model, however, was imprecise in terms of the extent to which exposure would be increased with estimates ranging from little effect to approximately eight to ten times the exposure in normal healthy volunteers. The study was performed using a low dose to enhance the understanding of the impact of moderate hepatic impairment. The lower dose was used to ensure exposure levels would remain well below the NOAEL.

Table 10.11: Pharmacokinetic Data from a Clinical Pharmacology Hepatic Impairment Trial

Subject	Group	AUC (ng.h/mL)	Cmax (ng/mL)
100	HEALTHY	2572	718
101	HEPATIC	2862	374
102	HEPATIC	5225	302
103	HEPATIC	3709	441
104	HEPATIC	3866	258
105	HEPATIC	2675	382
106	HEALTHY	2911	504
107	HEPATIC	4321	439
108	HEPATIC	5801	434
109	HEALTHY	2701	466
110	HEALTHY	2374	606
111	HEPATIC	3023	409
112	HEALTHY	3023	467
HEALTHY (No Liver Disease) HEPATIC (Moderate Liver Disease)			

Table 10.11: Pharmacokinetic Data from a Clinical Pharmacology Hepatic Impairment Trial

Subject	Group	AUC (ng.h/mL)	Cmax (ng/mL)
113	HEALTHY	2344	449
114	HEALTHY	2544	386
115	HEPATIC	3352	343
116	HEALTHY	2802	422
117	HEPATIC	2768	422
118	HEPATIC	2489	554
119	HEALTHY	3715	385
120	HEPATIC	3740	488
121	HEALTHY	2088	487
122	HEALTHY	2038	474
123	HEALTHY	1703	592
124	HEPATIC	2711	301
201	HEPATIC	3164	349
202	HEPATIC	1998	303
203	HEPATIC	4270	316
204	HEPATIC	5501	773
205	HEALTHY	1983	553
206	HEALTHY	3494	728
207	HEALTHY	3962	478
208	HEALTHY	3106	493
209	HEPATIC	2897	432
210	HEALTHY	1598	392
HEALTHY (No Liver Disease) HEPATIC (Moderate Liver Disease)			

Code to analyse such data is provided below and is very similar to that used in previous model validity exercises. In cases where data are collected from mild and severe liver impairment patients, the groupings may be changed to accommodate this, or the Child-Pugh scores themselves can be used to examine the correlation between scores and the pharmacokinetics. We will consider such an approach in the next section, using renal impairment data.

Hepatic Impairment AUC Assessment Example - SAS `proc mixed` *code:*

```
proc mixed method=reml data=liver;
    class subject group;
    model lnauc=group/s
    ddfm=kenwardroger cl alpha=.1;
    lsmeans group/cl alpha=0.1;
    repeated /group=group subject=subject;
    ods output LSMeans=auc;
    run;
```

For logAUC, the resulting back-transformed model estimates are presented in Table 10.12. AUC was increased (as expected) in the hepatic impaired patients. Assessment of Cmax is left as an exercise for interested readers, and may be performed using code on the website accompanying this book.

Table 10.12 *Estimated Population Parameters from Evaluation of logAUC as a Function of Group*

Group	Estimated Mean AUC	90% CI
HEALTHY	2653	2296, 2861
HEPATIC	3433	3045, 3869
HEALTHY (No Liver Disease) HEPATIC (Moderate Liver Disease)		

According to the suggestion in regulatory guidance [136], if a doubling in extent of exposure is observed relative to the levels used to achieve efficacy while maintaining safety in the normal patient population, the dose in hepatic-impaired patients should be adjusted downward. If exposures cannot be kept clear of the NOAEL, one would presumably not wish to expose patients to such a risk and might contraindicate. If desired, a no-effect claim may be established if a two one-sided test (similar to that used for average bioequivalence) with a clinically relevant threshold is set up *a priori* in the protocol [136], but we omit discussion of such an approach here as inference and labelling based on such trials most often utilizes expert clinical assessment of estimated model parameters without such formal statistical testing.

10.6 Kidney Disease

Most drugs are eliminated unchanged by the kidney or by metabolism in the liver [124]. As with hepatic impairment, renal impairment can be caused by a variety of factors, and we will not discuss these further here. As age increases, this also results in impaired functioning of the kidney.

For drugs which are eliminated from circulation by the kidney, impaired function is expected to result in decreased clearance [12]. Decreased clearance would be expected to result in increased exposure, and as with hepatic impairment, this may result in increased likelihood of adverse experiences.

Creatinine clearance (CLcr) is a parameter often used to describe renal function [124]. This endpoint may be derived as [124]:

$$CLcr = \frac{(140 - age(yrs))weight(kg)}{72(serum - creatinine(mg/dL))}.$$

This formula is multiplied by 0.85 for female subjects and represents steady-state renal function. Severity of impairment is typically characterized using these values as [124]:

1. Healthy (CLcr $>$ 80 mL/min),
2. Mild (CLcr from 50-80 mL/min),
3. Moderate (CLcr from 30-50 mL/min),
4. Severe (CLcr $<$ 30 mL/min), and
5. ESRD (requiring dialysis).

While building a population pharmacokinetic model (see Section 10.1), the impact of renal function (assessed using creatinine clearance) on estimated parameters for plasma clearance will generally be assessed. As with hepatic impairment if there is good scientific evidence to support this being minor (i.e., renal clearance plays only a small role in elimination and metabolism of the drug), then one need not study the issue further in drug development [124]. Note that this involves some degree of subjectivity; hence, in practice, a study is generally done to validate the understanding from the population pharmacokinetic model.

Design of such a trial is similar to the other validation exercises discussed in this chapter. Sample size is selected to provide appropriate precision in study findings in similar fashion to the approach used for age, gender, ethnicity, and hepatic trials, and further discussion (and code) is omitted here. Roughly equal numbers of subjects in each renal impairment severity class are recruited and given a single dose of drug with a typical clinical pharmacology pharmacokinetic sampling scheme performed. One may also study the ends of the impairment spectrum (severe versus healthy) before enrolling mild and moderates [124].

Often mentioned in the context of renal impairment is the importance of protein binding and consideration of (and derivation of) unbound concentrations and estimates of rate and extent of exposure. Drug molecules bound to protein in plasma are not active, and are often removed from circulation by the kidney. Drug not bound to protein is typically the active component which, reaching the site of action, is presumed to elicit a pharmacodynamic response in the body (see Chapters 1 and 2). As protein binding may be impacted by kidney function, typically one blood sample is collected in such studies for each subject to estimate the degree of drug protein binding. If the degree of binding is pronounced (greater than 80%), the unbound concentration is used to derive an estimate of unbound AUC (AUCu) and unbound Cmax (Cmaxu) by straightforward multiplication. An example data set may be found in Table 10.13. Note that in this experiment, 60 mL/min was used as the cut-off between mild and moderate renal impairment as it pre-dated the [124] guidance.

Table 10.13: Pharmacokinetic Data from a Clinical Pharmacology Renal Impairment Trial

Group	Subject	CLcr	AUC (ng.h/mL)	Cmax (ng/mL)	AUCu (ng.h/mL)	Cmaxu (ng/mL)
HEALTHY	107	105	1523	407	1.68	0.448
HEALTHY	116	87	2426	409	3.40	0.573
HEALTHY	117	92	3919	341	.	.
HEALTHY	126	105	3351	606	.	.
HEALTHY	127	90	1851	474	2.78	0.711
HEALTHY	128	101	3487	444	5.23	0.666
HEALTHY	130	84	3719	592	7.44	1.184
HEALTHY	131	82	3046	400	4.87	0.640
HEALTHY	138	96	3282	474	4.59	0.664
HEALTHY	215	94	2823	424	4.80	0.721
HEALTHY	218	81	2765	584	3.59	0.759
HEALTHY	219	100	1860	377	3.91	0.792
MILD	102	67	2635	392	4.48	0.666
MILD	110	72	2321	320	.	.
MILD	113	68	4498	440	7.65	0.748
MILD	115	68	2727	460	4.09	0.690
MILD	118	67	3226	681	4.52	0.953
MILD	121	66	2653	401	4.51	0.682
MILD	122	73	6710	458	11.41	0.779
MILD	123	69	3991	507	6.39	0.811
MILD	124	65	2304	347	3.23	0.486
MILD	207	64	3254	455	4.88	0.683
MILD	208	71	3364	670	4.37	0.871
MILD	210	61	2271	476	3.18	0.666
MILD	212	74	3137	500	6.59	1.050
MILD	216	64	1560	323	2.18	0.452
MILD	217	71	2235	374	3.80	0.636
MODERATE	105	33	2375	495	3.80	0.792
MODERATE	106	49	3658	389	5.85	0.622
MODERATE	108	44	6638	710	13.28	1.420
MODERATE	111	53	2167	427	3.03	0.598
MODERATE	112	48	3445	517	4.82	0.724
MODERATE	114	46	3670	565	6.61	1.017
MODERATE	120	57	3108	440	5.59	0.792
MODERATE	125	58	3959	599	6.33	0.958
MODERATE	132	55	2211	286	3.76	0.486
MODERATE	133	53	3138	442	5.02	0.707
MODERATE	134	54	3003	572	3.90	0.744
MODERATE	135	53	4187	469	.	.
MODERATE	202	51	2627	337	3.68	0.472
MODERATE	203	54	2718	474	3.81	0.664
MODERATE	204	55	3410	558	5.12	0.837
MODERATE	205	55	3314	405	4.97	0.608
MODERATE	209	58	2105	352	2.53	0.422
MODERATE	213	43	2520	504	3.53	0.706
SEVERE	101	15	2290	230	.	.
SEVERE	103	22	2825	262	.	.
SEVERE	104	17	2427	370	4.13	0.629

HEALTHY (No Renal Disease; CLcr> 80)
MILD (Mild Renal Disease; 60 <CLcr≤ 80)
MODERATE (Moderate Renal Disease; 30 <CLcr≤ 60)
SEVERE (Severe Renal Disease; CLcr≤ 30)

Table 10.13: Pharmacokinetic Data from a Clinical Pharmacology Renal Impairment Trial

Group	Subject	CLcr	AUC (ng.h/mL)	Cmax (ng/mL)	AUCu (ng.h/mL)	Cmaxu (ng/mL)
SEVERE	109	22	2704	527	4.06	0.791
SEVERE	119	23	2237	395	6.04	1.067
SEVERE	129	18	1490	233	2.53	0.396
SEVERE	136	27	1407	329	2.67	0.625
SEVERE	137	21	3415	447	6.83	0.894
SEVERE	201	24	2325	404	.	.
SEVERE	206	6	1675	259	5.36	0.829
SEVERE	211	19	1974	329	4.15	0.691
SEVERE	214	22	2705	526	7.03	1.368

HEALTHY (No Renal Disease; CLcr$>$ 80)
MILD (Mild Renal Disease; 60 $<$CLcr$\leq$ 80)
MODERATE (Moderate Renal Disease; 30 $<$CLcr$\leq$ 60)
SEVERE (Severe Renal Disease; CLcr$\leq$ 30)

Code to analyse such data is provided below and is very similar to that used in previous model validity exercises. Variability is allowed to change with group using the `repeated` statement, and the relationship of the pharmacokinetic endpoint of interest (in this example, AUC) is modelled on the logscale as a function of creatinine clearance. The `estimate` statements are used to output estimates of mean AUC at various levels of creatinine clearance.

Renal Impairment AUC Assessment Example - SAS `proc mixed` *code:*

```
proc mixed method=reml data=renal;
    class subject group;
    model lnauc=clcr/s
    ddfm=kenwardroger cl alpha=.1 outp=out;
    estimate '80' intercept 1 clcr 80/cl alpha=0.1;
    estimate '70' intercept 1 clcr 70/cl alpha=0.1;
    estimate '60' intercept 1 clcr 60/cl alpha=0.1;
    estimate '50' intercept 1 clcr 50/cl alpha=0.1;
    estimate '40' intercept 1 clcr 40/cl alpha=0.1;
    estimate '30' intercept 1 clcr 30/cl alpha=0.1;
    estimate '20' intercept 1 clcr 20/cl alpha=0.1;
    estimate '10' intercept 1 clcr 10/cl alpha=0.1;
    repeated /group=group subject=subject;
    ods output Estimates=outest;
    run;
```

For AUC, the resulting back-transformed model estimates are presented in Table 10.14. No relationship between creatinine clearance and AUC was observed in the renally impaired patients. Analysis of Cmax and unbound AUC and Cmax are left as an exercise for interested readers, and may be performed using code on the website accompanying this book.

Generally, a log-linear relationship of total and unbound AUC and Cmax with creatinine clearance is observed. If not, transformation of creatinine clearance using a power model generally suffices to adequately

Table 10.14 *Estimated Population Parameters from Evaluation of logAUC as a function of Creatinine Clearance*

Creatinine Clearance	Estimated Mean AUC	90% CI
80	2953	2651,3289
70	2869	2625,3136
60	2787	2580,3011
50	2708	2506,2926
40	2631	2406,2878
30	2557	2293,2851
20	2484	2177,2835
10	2413	2062,2825

describe the data. As such a model has already been described in the context of dose-proportionality (see Chapter 7), this is not discussed further here.

As with hepatic impairment, based on these findings, the population pharmacokinetic model may be rebuilt, if appropriate. Dose is typically adjusted in renally impaired patients to achieve concentrations that are expected to be safe and effective. Labelling statements based on data like those described above provide the basis for the selection of dose-adjustment or contraindication. [124].

10.7 Technical Appendix

10.7.1 Models, Derivations, and Software in Population Pharmacokinetic Models

Models such as

$$c_{it} = (e^{-k_{ei}t} - e^{-k_{ai}t})\frac{k_{ei}k_{ai}(Dose)}{Cl_i(k_{ai} - k_{ei})} + \varepsilon_{it} \tag{10.1}$$

may be used to easily derive estimates for Tmax, Cmax, and AUC. We provide one such example here based on the estimated parameters from the fitted model.

To derive Tmax, take the first derivative of $\hat{c}_{it}$ (the fitted model) with respect to t. The resulting equation is:

$$\frac{d\hat{c}_{it}}{dt} = \frac{\hat{k}_{ai}\hat{k}_{ei}(Dose)}{\hat{C}l_i(\hat{k}_{ai} - \hat{k}_{ei})}(-\hat{k}_{ei}e^{-\hat{k}_{ei}t} + \hat{k}_{ai}e^{-\hat{k}_{ai}t}).$$

Setting this equal to zero and solving for t yields an estimate for Tmax of

$$Tmax = \frac{\ln \hat{k}_{ai} - \ln \hat{k}_{ei}}{\hat{k}_{ai} - \hat{k}_{ei}}.$$

An estimate for Cmax may be derived by taking the predicted concentration at this time point:

$$Cmax = (e^{-\hat{k}_{ei}Tmax} - e^{-\hat{k}_{ai}Tmax})\frac{\hat{k}_{ei}\hat{k}_{ai}(Dose)}{\hat{C}l_i(\hat{k}_{ai} - \hat{k}_{ei})}.$$

To derive $AUC(0-\infty)$, take the integral from zero to infinity of $\hat{c}_{it}$ with respect to time (t):

$$\int_0^\infty \hat{c}_{it}dt = \frac{\hat{k}_{ai}\hat{k}_{ei}(Dose)}{\hat{C}l_i(\hat{k}_{ai} - \hat{k}_{ei})}\int_0^\infty (e^{-\hat{k}_{ei}t} - e^{-\hat{k}_{ai}t})dt.$$

Integration yields:

$$\frac{\hat{k}_{ai}\hat{k}_{ei}(Dose)}{\hat{C}l_i(\hat{k}_{ai} - \hat{k}_{ei})}(\frac{1}{\hat{k}_{ei}} - \frac{1}{\hat{k}_{ai}}) = \frac{Dose}{\hat{C}l_i}.$$

As stated in Chapter 7, we chose here to utilize SAS for the nonlinear mixed effect modelling of data; however, several other statistical packages are readily available (SPLUS, NONMEM, WINNONLIN, P-KBUGS, etc., [365]) and may be used for this purpose. Readers interested in more details of these software packages should see [468] and [435].

10.7.2 Determining Precision for Absolute and Relative Bioavailability Studies

The approach to determine precision in estimates of absolute bioavailability $\hat{F}$ or to the ratio of relative bioavailability for differing formulations administered by the same route is similar to that shown for drug-interaction trials in Chapter 7. SAS code is as follows. Use of a randomized 2 × 2 cross-over is assumed, and the sample size (`n`) and standard deviation of AUC (`sigmaW`) should be entered by the user.

Sample Size Code for Precision in Absolute Bioavailability and Relative Bioavailability Studies:

```
data a;
    * total number of subjects
    (needs to be a multiple of number
    of two);
n=20; seq=2;
    * significance level;
a=0.05;
    * variance of difference of two observations
    on the log scale;
    * sigmaW = within-subjects standard deviation;
sigmaW=0.2; s=sqrt(2)*sigmaW;
    * error degrees of freedom for cross-over
    with n subjects in total
    assigned equally to seq sequences;
n2=n-seq;
    run;

data b; set a;
* calculate precision;
    t=tinv(1-a,n2);
    SE=s/(sqrt(n));
* precision on log-scale;
    w=t*SE;
* precision on natural-scale;
    exp_w=(exp(t*SE)-1)*100;
    run;

proc print data=b; run;
```

10.7.3 Determining Precision for Pediatric, Elderly, and Gender Pharmacokinetic Studies

The approach to determine precision in estimates of age and gender trials is similar to that shown for drug-interaction trials in Chapter 7. SAS code is as follows. Use of a non-randomized parallel-group trial is assumed, and the sample size (`n`) and between-subject standard deviation of AUC (`sigma`) should be entered by the user.

Sample Size Code for Precision in Age and Gender Studies:

```
data a;
    * number of subjects per grouping;
    n=10;
    * significance level;
    a=0.05;
    * sigma = standard deviation;
    sigma=0.3; s=sigma;
    * error degrees of freedom for each group;
    n2=n-1;
    run;

data b; set a;
* calculate precision;
    t=tinv(1-a,n2);
    SE=s/(sqrt(n));
* precision on log-scale;
    w=t*SE;
* precision on natural-scale;
* plus or minus (percentage) of mean exposure;
    exp_w=(exp(t*SE)-1)*100;
    run;

proc print data=b; run;
```

CHAPTER 11

Epilogue

Many people in business and medicine regard statistics as at best a nuisance, and at worst a hinderance to science. For example, Einstein liked to say that 'God doesn't gamble.' [54]. That is likely true in the long run (in statistical terms, in the limit), but in the short term while making drugs, we cannot operate with complete certainty and have to depend upon statistics to guide us in making safe, effective, quality products. In clinical drug development, statistics are used to quantify the uncertainty associated with human use of drugs - not to eliminate uncertainty. In a perfect world, it would be perfectly clear whether to use a drug or not (the drug is either safe, effective, and made well or it is not), but in practice, statistics are used to measure the outcome of studies used as tools to assess these properties. If they are not used well, the trend toward increased length and cost in drug development [137] are very likely to continue.

Clinical pharmacology and many aspects of drug development are evolving, and will continue to do so. These changes are good as they would be expected to improve the drugs that are produced for the people who need them. Statistically, changes such as these represent new challenges, but the raw materials to meet the needs of the science are available. Change is not so bad once you get used to it.

The future of Statistics in Clinical Pharmacology lies in ***learning*** *(not confirming). Design of studies and bioequivalence analyses can be automated by sponsoring companies themselves or by commercial software companies. Indeed, some software companies already claim to do so, and it is to be expected that more will appear in the future. This frees up statisticians to spend more time on the parts of drug development that are in need of attention. Confirmatory work like bioequivalence testing should soon no longer be an activity directly involving statisticians but will be handed off to clinical scientists and pharmacokineticists with statisticians only being consulted as needed.*

To reiterate, the future of statistics in clinical pharmacology lies in other areas - in particular, better understanding of quantitative aspects of safety and efficacy assessment in Phases I and IIa. Better control of these should lead to less Phase III portfolio attrition. Particularly, safety in the use of drug products could use some quantitative enhancements.

We hope you have found this book useful in making some of the concepts associated with statistics in clinical pharmacology more transparent, and we hope it has provided practical tools for people working in this area of drug development. We wish all our readers good luck with the application of the principles described in this book and look forward to many exciting discoveries in clinical pharmacology and statistics in the coming years.

Bibliography

[1] Aarons, L., Karlsson, M., Mentre, F., Rombout, F., Steimer, J.-L., van Peer, A., and Invited COST B15 Experts (2001) The role of modelling and simulation in phase I drug development. *European Journal of Pharmaceutical Sciences*, **13**, 115–122.

[2] Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 267–281.

[3] Altman, D.G., Bland, J.M. (1983) Measurement in medicine: the analysis of method comparison studies. *The Statistician*, **32**, 307–317.

[4] Altman, D.G., Bland, J.M. (1995) Absence of evidence is not evidence of absence. *British Medical Journal*, **311**, 485.

[5] Amankwa, A., Krishnan, S., Tisdale, J. (2004) Torsades de pointes associated with fluoroquinilones: importance of concommitant risk factors. *Clinical Pharmacology and Therapeutics*, **75(3)**, 242–7.

[6] Anderson, S. (1993) Individual bioequivalence: a problem of switchability [with discussion]. *Biopharmaceutical Reports*, **2**, 1–11.

[7] Anderson, S. (1995) Current issues in individual bioequivalence. *Drug Information Journal*, **29**, 961–964.

[8] Anderson, S., Hauck W.W. (1983) A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistical Theory and Methods*, **12**, 2663–2692.

[9] Anderson, S., Hauck W.W. (1990) Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, **18**, 259–273.

[10] Anderson, S., Hauck W.W. (1996) The transitivity of bioequivalence testing: potential for drift. *International Journal of Clinical Pharmacology and Therapeutics*, **34**, 369–374.

[11] Ansbacher, R. (1990) Interchangeability of low dose oral contraceptives. *Contraception*, **43**, 139–147.

[12] Atkinson, A., Daniels, C., Dedrick, R., Grudzinskas, C., Markey, S., eds. (2001) *Principles of Clinical Pharmacology.* Academic Press, San Diego.

[13] Australia Therapeutic Goods Administration, Australian Regulatory Guidelines for Prescription Medicines (2002) Appendix 15, Biopharmaceutical Studies.

[14] Balakrishnan, N., Ma, C.W. (1990) A comparative study of various tests for the equality of two population variances. *Journal of Statistical Computing and Simulation*, **35**, 41–89.

[15] Balthasar, J.P. (1999) Bioequivalence and bioequivalency testing. *American Journal of Pharmaceutical Education*, **63**, 194–198.

[16] Barbieri M.M., Liseo, B., Petrella, L. (2000) Bayes factors for Fieller's problem. *Biometrika*, **87**, 717–723.

[17] Barrett, J.S., Batra, V., Chow, A., Cook, J., Gould, A.L., Heller, A., Lo, M.-W., Patterson, S.D., Smith, B.P., Stritar, J.A., Vega, J.M., Zariffa, N. (2000) PhRMA perspective on population and individual bioequivalence and update to the PhRMA perspective on population and individual bioequivalence. *Journal of Clinical Pharmacology*, **40**, 561–572.

[18] Bartlett, M.S. (1936) Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, Series A*, **154**, 124–137.

[19] Basson, R.P., Ghosh, A., Cerimele, B.J., DeSante, K.A., Howey, D.C. (1998) Why rate of absorption inferences in single dose bioequivalence studies are often inappropriate. *Pharmaceutical Research*, **15**, 276–279.

[20] Bauer, P. (1991) Multiple testing in clinical trials. *Statistics in Medicine*, **10**, 871–890.

[21] Bauer, P., Bauer, M.M. (1994) Testing equivalence simultaneously for location and dispersion of two normally distributed populations. *Biometrical Journal*, **6**, 643–660.

[22] Bauer, P., Kieser, M. (1996) A unifying approach for confidence intervals and testing of equivalence and difference. *Biometrika*, **83**, 934–937.

[23] Bazett, H. (1920) An analysis of time relations of electrocardiograms. *Heart*, **7**, 353–367.

[24] Bekersky, I., Dressler, D., Colburn, W., Mekki, Q. (1999) Bioequivalence of 1 and 5 mg Tacrolimus capsules using a replicate study design. *Journal of Clinical Pharmacology*, **39**, 1032–1037.

[25] Bellavance, F., Tardif, S. (1995) A nonparametric approach to the analysis of three-treatment three-period cross-over data. *Biometrika*, **82**, 865–875.

[26] Benet, L.Z., Goyan, J.E. (1995) Bioequivalence and narrow therapeutic index drugs. *Pharmacotherapy*, **15**, 433–440.

[27] Benet, L.Z. (1999) Understanding bioequivalence testing. *Transplantation Proceedings*, **31, Suppl A**, 7S–9S.

[28] Berger, R.L. (1992) Multiparametric hypothesis testing and acceptance sampling. *Technometrics*, **24**, 295–300.

[29] Berger, R.L., Hsu, J.C. (1996) Bioequivalence trials, intersection–union tests, and equivalence confidence sets. *Statistical Science*, **11**, 283–319.

[30] Bhoj, D.S. (1979) Testing equality of variances of correlated variates with incomplete data on both responses. *Biometrika*, **66**, 681–683.

[31] Bickel, P.J., Doksum, K.A. (1977) *Mathematical Statistics*. Holden Day, San Francisco.

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

[33] Blackwelder, W.C. (1982) Proving the null hypothesis in clinical trials. *Controlled Clinical Trials*, **3**, 345–353.

[34] Blume, H.H., Midha, K.K., eds. (1993) Conference Report. In: Bio-International: Bioavailability, Bioequivalence, and Pharmacokinetics. *Medpharm Stuttgart*, 13–23.

[35] Boddy, A.W., Snikeris, F.C., Kringle, R.O., Wi, G.C.-G., Oppermann, J.A., Midha, K.K. (1995) An approach for widening the bioequivalence acceptance limits in the case of highly variable drugs. *Pharmaceutical Research*, **12**, 1865–1868.

[36] Bois, F.Y., Tozer, T.N., Hauck, W.W., Chen, M.L., Patnaik, R., Williams, R. (1994) Bioequivalence: Performance of several measures of extent of absorption. *Pharmaceutical Research*, **11**, 715–722.

[37] Bonate, P., Howard, D., eds. (2004) *Pharmacokinetics in Drug Development: Clinical Study Design and Analysis*. AAPS Press, Arlington, VA.

[38] Box, G.E., Cox, D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211–243.

[39] Box, G.E. (1966) Use and abuse of regression. *Technometrics*, **8**, 625–629.

[40] Box, G.E. (1979) Robustness in the strategy of scientific model building. In *Robustness in Statistics*, Launer, R.L., Wilkinson, G.N., eds., 201–236. Academic Press: New York.

[41] Boxtel, C., Holford, N., Danhof, M., eds. (1992) *The In Vivo Study of Drug Action*. Elsevier, NY.

[42] Breslow, N. with discussion (1990) Biostatistics and Bayes. *Statistical Science*, **5**, 269–298.

[43] Bristol, D.R. (1991a) Testing equality of treatment variances in a 2×2 cross-over study. *Journal of Biopharmaceutical Statistics*, **1**, 185–192.

[44] Bristol, D.R. (1991b) A confidence interval for the ratio of treatment variances in a 2×2 cross-over study. *Journal of Biopharmaceutical Statistics*, **1**, 237–245.

[45] Brown, B.W. (1980) The cross-over experiment for clinical trials. *Biometrics*, **36**, 69–79.

[46] Brown, E.B., Iyer, H.K., Wang, C.M. (1997) Tolerance intervals for assessing individual bioequivalence. *Statistics in Medicine*, **16**, 803–820.

[47] Brown, H.K., Kempton, R.A. (1994) The application of REML in clinical trials. *Statistics in Medicine*, **13**, 1601–1617.

[48] Brown, L.D., Hwang, J.T.G., Munk, A. (1997) An unbiased test for the bioequivalence problem. *Annals of Statistics*, **25**, 2345–2367.

[49] Brown, L.B., Casella, G., Hwang, G. (1995) Optimal confidence sets, bioequivalence, and the limacon of Pascal. *Journal of the American Statistical Association*, **90**, 880–889.

[50] Brown, M.B., Forsythe, A.B. (1974) Robust tests for the equality of variances. *Journal of the American Statistical Association*, **69**, 364–367.

[51] Buice, R.G., Subramanian, V.S., Duchin, K.L., Uko-Nne, S. (1996) Bioequivalence of a highly variable drug: An experience with Nadolol. *Pharmaceutical Research*, **13**, 1109–1115.

[52] Burdick, R.K., Graybill, F.A. (1992) *Confidence Intervals on Variance Components*. Marcel Dekker, NY.

[53] Burdick, R.K., Sielken, R.L. (1978) Exact confidence intervals for linear combinations of variance components in nested classifications. *Journal of the American Statistical Association*, **73**, 632–635.

[54] Burnham, R. (2005) The man who remade the universe. *Astronomy*, **33**, 39–41.

[55] Burzykowski, T., Molenberghs, G., Buyse, M., eds. (2005) *The Evaluation of Surrogate Endpoints* Springer, NY.

[56] Calvert, R.T. (1996) Bioequivalence and generic prescribing: a pharmacy view. *Journal of Pharmacy and Pharmacology*, **48**, 9–10.

[57] Canadian Guidance for Industry (1992) Conduct and Analysis of Bioequivalence Studies - Part A.

[58] Canadian Guidance for Industry (1996) Conduct and Analysis of Bioequivalence Studies - Part B.

[59] Canafax, D.M., Irish, W.D., Moran, H.B., Squiers, E., Levy, R., Pouletty, P., First, M.R., Christians, U. (1999) An individual bioequivalence approach to compare the intra-subject variability of two Ciclosporin formulations, SangCya and Neoral. *Pharmacology*, **59**, 78–88.

[60] Carter, B.L., Noyes, M.A., Demmler, R.W. (1993) Differences in serum concentrations of and responses to generic verapamil in the elderly. *Pharmacotherapy*, **13**, 359–368.

[61] Cartwright, A.C. (1991) International harmonisation and consensus DIA meeting on bioavailability and bioequivalence requirements and standards. *Drug Information Journal*, **25**, 471–482.

[62] Casella, G., George, E.J. (1992) Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.

[63] Cerutti, R., Rivolta, G., Cavalieri, L., Di Giulio, C., Grossi, E., Vago, T., Baldi, G., Righini, V., Marzo, A. (1999) Bioequivalence of Levothyroxine tablets administered to a target population in steady state. *Pharmacological Research*, **39**, 193–201.

[64] Chaturvedi, A., Bhatti, M.I., Kumar, K. (2000) Bayesian analysis of disturbances variance in the linear regression model under assymetric loss assumptions. *Applied Mathematics and Computation*, **114**, 149–153.

[65] Chen, K-W., Li, G., Sun, Y. (1996) A confidence region approach for assessing equivalence in variability of bioavailability. *Biometrical Journal*, **4**, 475–487.

[66] Chen, M.-L. (1997) Individual bioequivalence - a regulatory update. *Journal of Biopharmaceutical Statistics*, **7**, 5–11.

[67] Chen, M.-L., Patnaik, R., Hauck, W.H., Schuirmann, D.J., Hyslop, T., Williams, R. (2000a) An individual bioequivalence criterion: regulatory considerations. *Statistics in Medicine*, **19**, 2821–2842.

[68] Chen, M.-L., Lee, S-C., Ng, M-J., Schuirmann, D., Lesko, L.J., Williams, R.L. (2000b) Pharmacokinetic analysis of bioequivalence trials: Implications for sex-related issues in clinical pharmacology and biopharmaceutics. *Clinical Pharmacology and Therapeutics*, **68**, 510–521.

[69] Chi, E.M. (1994) M-estimation in cross-over trials. *Biometrics*, **50**, 486–493.

[70] China State Drug Administration (2003) Drug Registration Regulation.

[71] Chinchilli, V.M. (1996) The assessment of individual and population bioequivalence. *Journal of Biopharmaceutical Statistics*, **6**, 1–14.

[72] Chinchilli, V.M., Elswick, R.K. (1997) The multivariate assessment of bioequivalence. *Journal of Biopharmaceutical Statistics*, **7**, 113–123.

[73] Chinchilli, V.M., Esinhart, J.D. (1996) Design and analysis of intra-subject variability in cross-over experiments. *Statistics in Medicine*, **15**, 1619–1634.

[74] Chow, S.C., Liu, J. (2000) *Design and Analysis of Bioavailability and Bioequivalence Studies, 2ed.* Marcel Dekker, New York.

[75] Chow, S.C., Shao, J. (1988) A new procedure for the estimation of variance components. *Statistics and Probability Letters*, **6**, 349–355.

[76] Chow, S.C., Shao, J., Wang, H. (2002a) Individual bioequivalence testing under 2 × 3 designs. *Statistics in Medicine*, **21**, 629–48.

[77] Chow, S.C., Shao, J., Hu, O. (2002b) Assessing sensitivity and similarity in bridging studies. *Journal of Biopharmaceutical Statistics*, **12**, 385–400.

[78] Colaizzi, J.L., Lowenthal, D.T. (1986) Critical therapeutic categories: A contraindication to generic substitution? *Clinical Therapeutics*, **8**, 370–379.

[79] Colburn, W.A., Keefe, D.L. (2000) Bioavailability and bioequivalence: Average, population, and/or individual. *Journal of Clinical Pharmacology*, **40**, 559–600.

[80] Conforti, M., Hochberg, Y. (1987) Sequentially rejective pairwise testing procedures. *Journal of Statistical Planning and Inference*, **17**, 193–208.

[81] Cornell, R.G. (1980) Evaluation of bioavailability data using nonparametric statistics. *In Drug Absorption and Disposition: Statistical Considerations; Albert, K., ed.*, 51–57: American Pharmaceutical Association.

[82] Cornell, R.G. (1991) Nonparametric tests of dispersion for the two period cross-over design. *Commumications in Statistical Theory and Methodology*, **20**, 1099–1106.

[83] Cox, D.R. (1967) Fieller's theorem and a generalization. *Biometrika*, **54**, 567–572.

[84] Crowder, M.J., Kimber, A.C., Smith, R.L., Sweeting, T.J. (1991) *Statistical Analysis of Reliability Data.*, Chapman and Hall, London.

[85] Crow, E.L., Shimizu, K. (1988) *Lognormal Distributions.* Marcel Dekker, New York.

[86] Cytel (1995) *StatXact 3 for Windows: Statistical Software for Exact Nonparametric Inference (User Manual).* Cytel Software Corporation, Cambridge, MA.

[87] D'Angelo, G., Potvin, D., Turgeon, J. (2001) Carry-over effects in bioequivalence studies. *Journal of Biopharmaceutical Statistics*, **11**, 27–36.

[88] Davidian, M., Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data.* Chapman and Hall, London.

[89] Diletti, E., Hauschke, D., Steinijans, V.W. (1991) Sample size determination for bioequivalence assessment by means of confidence intervals. *International Journal of Clinical Pharmacology, Therapeutics, and Toxicology*, **29**, 1–8.

[90] DiMasi, J. (2001) New drug development in the USA from 1963 to 1999; Risks in new drug development - Approval success rates for investigational drugs. *Clinical Pharmacology and Therapeutics*, **69**, 286–307.

[91] Dmitrienko, A., Smith, B. (2003) Repeated-measures models in the analysis of QT-interval data. *Pharmaceutical Statistics*, **2**, 175–190.

[92] Dragalin, V., Fedorov, V. (1999a) Kullback–Liebler Distance for evaluating bioequivalence. *SB BDS Technical Report 1999-03.*

[93] Dragalin, V., Fedorov, V. (1999b) The Total Least Squares Method in Individual Bioequivalence Evaluation. *SB BDS Technical Report 1999-04.*

[94] Dragalin, V., Fedorov, V. (2004) Adaptive Model-Based Designs for Dose-Finding Studies. *GSK BDS Technical Report 2004-02.*

[95] Dunnett, C.W., Gent, M. (1977) Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. *Biometrics*, **33**, 593–602.

[96] Efron, B., Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Chapman and Hall, New York.

[97] Ekbohm, E. (1981) A test for the equality of variances in the paired case with incomplete data. *Biometrical Journal*, **3**, 261–265.

[98] Ekbohm, G., Melander, H. (1989) The subject-by-formulation interaction as a criterion of interchangeability of drugs. *Biometrics*, **45**, 1249–1254.

[99] El-Tahtawy, A.A., Tozer, T.N., Harrison, F., Lesko, L., Williams, R. (1998) Evaluation of bioequivalence of highly variable drug using clinical trial simulations. II: Comparison of Single and Multiple-Dose Trials using AUC and Cmax. *Pharmaceutical Research*, **15**, 98–104.

[100] EMEA, European Agency for the Evaluation of Medicinal Products, Committee for Proprietary Medicinal Products (2001) Guidance on the Investigation of Bioavailability and Bioequivalence. http://www.emea.eu.int/index/

[101] EMEA, European Agency for the Evaluation of Medicinal Products, Committee for Proprietary Medicinal Products (2004) Note for Guidance on Clinical Investigation of Medicinal Products in the Treatment of Lipid Disorders. http://www.emea.eu.int/index/

[102] Endrenyi, L., Fritsch, S., Yan, W. (1991) Cmax/AUC is a clearer measure than Cmax for absorption rates in investigations of bioequivalence. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, **29**, 394–399.

[103] Endrenyi, L., Schulz, M. (1993) Individual variation and the acceptance of average bioequivalence. *Drug Information Journal*, **27**, 195–201.

[104] Endrenyi, L. (1994) A method for the evaluation of individual bioequivalence. *International Journal of Clinical Pharmacology and Therapeutics*, **32**, 497–508.

[105] Endrenyi, L., Czimadia, F., Tothfalusi, L, Chen, M.L. (1998a) Metrics comparing simulated early concentration profiles for the determination of bioequivalence. *Pharmaceutical Research*, **15**, 1292–1299.

[106] Endrenyi, L., Amidon, G.L., Midha, K.K., Skelly, J.P. (1998b) Individual bioequivalence: Attractive in principle, difficult in practice. *Pharmaceutical Research*, **15**, 1321–1325.

[107] Endrenyi, L., Hao, Y. (1998c) Assymetry of the mean-variability tradeoff raises questions about the model in investigations of individual bioequivalence. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, **36**, 1–8.

[108] Endrenyi, L., Midha, K.K. (1998d) Individual bioequivalence–has its time come? *European Journal of Pharmaceutical Sciences*, **6**, 271–277.

[109] Endrenyi, L., Tothfalusi, L. (1999) Subject-by-formulation interaction in determination of individual bioequivalence: Bias and prevalence. *Pharmaceutical Research*, **16**, 186–190.

[110] Endrenyi, L., Taback, N., Tothfalusi, L. (2000) Properties of the estimated variance component for subject-by-formulation interaction in studies of individual bioequivalence. *Statistics in Medicine*, **19**, 2867–2878.

[111] Ereshefsky, L., Meyer, M. (2001) Comparison of the bioequivalence of generic versus branded Clozapine. *Journal of Clinical Psychiatry*, **62(Suppl 5)**, 3–27.

[112] Esinhart, J.D., Chinchilli, V.M. (1994a) Extensions to the use of tolerance intervals for the assessment of individual bioequivalence. *Journal of Biopharmaceutical Statistics*, **4**, 39–52.

[113] Esinhart, J.D., Chinchilli, V.M. (1994b) Sample size considerations for assessing individual bioequivalence based on the method of tolerance intervals. *International Journal of Clinical Pharmacology and Therapeutics*, **32**, 26–32.

[114] FDA Guidance (1987) Guidance for In-Vivo Bioequivalence Study for Slow Release Potassium-Chloride Tablets and Capsules. http://www.fda.gov/cder/guidance/

[115] FDA Guidance (1989) Guidance for the In-Vitro Portion of Bioequivalence Requirements for Metaproterenol Sulfate and Albuterol Inhalation Aerosols (Metered Dose Inhalers). http://www.fda.gov/cder/guidance/

[116] FDA Guidance (1992) Statistical Procedures for Bioequivalence Studies Using a Standard Two Treatment Cross-over Design.

[117] FDA Guidance (1993) Interim Guidance - Cholestyramine Powder In-Vitro Bioequivalence. http://www.fda.gov/cder/guidance/

[118] FDA Guidance (1994) Phenytoin Sodium Capsules, Tablets, and Suspension In VIvo Bioequivalence and In-Vitro Dissolution Testing. http://www.fda.gov/cder/guidance/

[119] FDA Guidance (1995a) Topical Dermatologic Corticosteroids: In-Vivo Bioequivalence. http://www.fda.gov/cder/guidance/

[120] FDA Guidance (1995b) Content and Format of Investigational New Drug Applications (INDs) for Phase 1 Studies of Drugs, Including Well-Characterised, Therapeutic, Biotechnology-derived Products. http://www.fda.gov/cder/guidance/

[121] FDA Guidance (1996) Clozapine Tablets In-Vivo Bioequivalence and In-Vitro Dissolution Testing. http://www.fda.gov/cder/guidance/

[122] FDA Preliminary Draft Guidance (1997a) In Vivo Bioequivalence Studies Based on Population and Individual Bioequivalence Approaches.

[123] FDA Guidance (1997b) Drug Metabolism, Drug Interaction Studies in the Drug Development Process: Studies In-vitro. http://www.fda.gov/cder/guidance/

[124] FDA Guidance (1998) Pharmacokinetics in Patients with Impaired Renal Function - Study Design, Data Analysis, and Recommendations for Dosing and Labeling. http://www.fda.gov/cder/guidance/

[125] FDA Draft Guidance (1999a) BA and BE Studies for Orally Administered Drug Products: General Considerations.

[126] FDA Draft Guidance (1999b) Average, Population, and Individual Approaches to Establishing Bioequivalence.

[127] FDA Guidance (1999c) In-Vivo Drug Metabolism, Drug-Interaction Studies - Study Design, Data Analysis, and Recommendations for Dosing and Labeling. http://www.fda.gov/cder/guidance/

[128] FDA Guidance (1999d) Population Pharmacokinetics. http://www.fda.gov/cder/guidance/

[129] FDA Guidance (2000a) Waiver of In Vivo Bioavailability and Bioequivalence Studies for Immediate-Release Solid Oral Dosage Forms Based on a Biopharmaceutics Classification System. http://www.fda.gov/cder/guidance/

[130] FDA Guidance (2000b) Bioavailability and Bioequivalence Studies for Orally Administered Drug Products: General Considerations.

[131] FDA Guidance (2001) Statistical Approaches to Establishing Bioequivalence. http://www.fda.gov/cder/guidance/

[132] FDA Guidance (2002a) Food-Effect Bioavailability and Fed Bioequivalence Studies. http://www.fda.gov/cder/guidance/

[133] FDA Draft Guidance (2002b) Estimating the Safe Starting Dose in Clinical Trials for Therapeutics in Adult Healthy Volunteers. http://www.fda.gov/cder/guidance/

[134] FDA Guidance (2003a) Exposure-Response Relationships – Study Design, Data Analysis, and Regulatory Applications. http://www.fda.gov/cder/guidance/

[135] FDA Guidance (2003b) Bioavailability and Bioequivalence Studies for Orally Administered Drug Products: General Considerations. http://www.fda.gov/cder/guidance/

[136] FDA Guidance (2003c) Pharmacokinetics in Patients with Impaired Hepatic Function: Study Design, Data Analysis, and Impact on Dosing and Labeling. http://www.fda.gov/cder/guidance/

[137] FDA Position Paper (2004a) Challenge and Opportunity on the Critical Path to New Medical Products.

[138] FDA Draft Guidance (2004b) Pharmacokinetics in Pregnancy - Study Design, Data Analysis, and Impact on Dosing and Labeling. http://www.fda.gov/cder/guidance/

[139] Guidance for Industry (2004c) E5 - Ethnic Factors in the Acceptability of Foreign Clinical Data: Questions and Answers. http://www.fda.gov/cder/guidance/

[140] Federer, W.T., Raghavarao, D. (1975) On augmented designs. *Biometrics*, **31**, 29–35.

[141] Fenech, A., Harville, D. (1991) Exact confidence sets for variance components in unbalanced mixed linear models. *The Annals of Statistics*, **19**, 1771–1785.

[142] Fieller, E. (1954) Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B*, **16**, 175–185.

[143] Fisher, L.D., van Belle, G. (1993) *Biostatistics, A Methodology for the Health Sciences*. John Wiley and Sons, New York.

[144] Fleiss, J.L. (1971) On the distribution of a linear combination of independent chi squares. *Journal of the American Statistical Association*, **66**, 142–144.

[145] Fleiss, J.L. (1986) Letter to the editor: On multiperiod cross-over studies. *Biometrics*, 449–450.

[146] Fleiss, J.L. (1989) A critique of recent research on the two treatment cross-over design. *Controlled Clinical Trials*, **10**, 237–243.

[147] Fleming, T., DeMets, D. (1996) Surrogate endpoints in clinical trials: Are we being misled?. *Annals of Internal Medicine*, **125**, 605–613.

[148] Fluehler, H., Hirtz, J., Moser, H.A. (1981) An aid to decision making in bioequivalence assessment. *Journal of Pharmacokinetics and Biopharmaceutics*, **9**, 235–243.

[149] Fridericia, L. (1920) Die systolendauer im elecktrokardiogramm bei normalen menschen und bei herzkranken. *Acta Medica Scandinavia*, **53**, 469–486.

[150] Freeman, P. (1989) The performance of the two-stage analysis of two treatment, two period cross-over trials. *Statistics in Medicine*, **8**, 1421–1432.

[151] Friedman, L., Furberg, C., DeMets, D. (1998) *Fundamentals of Clinical Trials, 3ed.* Springer, New York.

[152] Friesen, M.H., Walker, S.E. (1999) Are the current bioequivalence standards sufficient for the acceptance of narrow therapeutic index drugs? Utilization of a computer simulated warfarin bioequivalence model. *J Pharm Pharmaceut Sci*, **2**, 15–22.

[153] Gaffney, M. (1992) Variance components in comparative bioavailability studies. *Journal of Pharmaceutical Sciences*, **81**, 315–317.

[154] Garthwaite, P., Kadane, J., O'Hagan, A. (2005) Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, **100**, 680–700.

[155] Geischke, R., Steimer, J.-L. (2000) Pharmacometrics: modeling and simulation tools to improve decision making in clinical drug development. *European Journal of Drug Metabolism and Pharmacokinetics*, **25**, 49–58.

[156] Gelfland, A.E., Hills, S.E., Racine-Poon, A., Smith, A.F-M. (1990a) Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**, 972–985.

[157] Gelfand, A.E., Smith, A.F-M. (1990b) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

[158] Ghosh, K., Hesney, M., Sarkar, S., Thiyagarajan, B. (1996) Comprehensive criteria for population and individual bioequivalence. *Proceedings of the American Statistical Association, Biopharmaceutics Chapter*, 77-81.

[159] Ghosh, P., Khattree, R. (2003) Bayesian approach to average bioequivalence using Bayes' factor. *Journal of Biopharmaceutical Statistics*, **13**, 719–734.

[160] Giesbrecht, F., Burns, J. (1985) Two-stage analysis based on a mixed model: Large sample asymptotic theory and small-sample simulation results. *Biometrics*, **41**, 477–486.

[161] Gilks, W.R., Richardson, S., Spiegelhalter, D. (1996) *Markov Chain Monte Carlo in Practice*, Chapman and Hall, NY.

[162] Gobburu, J., Marroum, P. (2001) Utilisation of pharmacokinetic-pharmacodynamic modelling and simulation in regulatory decision making. *Clinical Pharmacokinetics*, **40**, 883–892.

[163] Godbillon, J., Cardot, J.B., LeCaillon, G., Sioufi, A. (1996) Bioequivalence assessment: a pharmaceutical industry perspective. *European Journal of Drug Metabolism and Pharmacokinetics*, **21**, 153–158.

[164] Goodman, S. (1994) The use of predictive confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, **121**, 200–206.

[165] Gould, A.L. (1995) Group sequential extensions of a standard bioequivalence testing procedure. *Journal of Pharmacokinetics and Biopharmaceutics*, **23**, 57–86.

[166] Gould, A.L. (1997) Discussion of individual bioequivalence by M.L. Chen. *Journal of Biopharmaceutical Statistics*, **7**, 23–29.

[167] Gould, A.L. (2000) A practical approach for evaluating population and individual bioequivalence. *Statistics in Medicine*, **19**, 2721–2740.

[168] Grahnan, A., Hammarlund, M., Lundqvist, T. (1984) Implications of intraindividual variability in bioavailability studies of furosemide. *European Journal of Clinical Pharmacology*, **27**, 595–602.

[169] Graybill, F.A., Wang, C.H. (1980) Confidence intervals on non-negative linear combinations of variances. *Journal of the American Statistical Association*, **75**, 869–873.

[170] Grieve, A.P. (1985) A Bayesian analysis of the two-period cross-over design for clinical trials. *Biometrics*, **21**, 467–480.

[171] Grieve, A.P. (1990) Cross-over vs. Parallel Designs. *In Statistical Methodology in the Pharmaceutical Sciences*, Berry, D., ed., 239–270, Marcel Dekker, New York.

[172] Grizzle, J.E. (1965) The two-period changeover design and its use in clinical trials. *Biometrics*, **21**, 467–480.

[173] Guilbaud, O. (1993) Exact inference about the within-subject variability in 2 × 2 cross-over studies. *Journal of the American Statistical Association*, **88**, 939–946.

[174] Guilbaud, O. (1999) Exact comparisons of means and variances in 2 × 2 cross-over trials. *Drug Information Journal*, **33**, 455–469.

[175] Hald, A. (1990) *A History of Probability and Statistics and Their Applications before 1750.* John Wiley and Sons, New York.

[176] Hald, A. (1998) *A History of Mathematical Statistics from 1750 to 1930.* John Wiley and Sons, New York.

[177] Hale, M., Gillespie, W.R., Gupta, S.K., Tuk, B., Holford, N. (1996) Clinical trial simulation. *Applied Clinical Trials*,**6**, 35–40.

[178] Hare, D., Foster, T. (1990) The Orange Book: the FDA's advice on therapeutic equivalence. *American Pharmacy*, **NS30**, 35–37.

[179] Harrell, F. (2001) *Regression Modelling Strategies.* Springer, New York.

[180] Harville, D.A. (1976) Confidence intervals and sets for linear combinations of fixed and random effects. *Biometrics*, **32**, 403–407.

[181] Harville, D.A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–340.

[182] Harville, D.A. (1990) BLUP (best linear unbiased prediction) and beyond. *AdStMetGen*, 239–276.

[183] Harville, D.A., Carriquiry, A.L. (1992) Classical and Bayesian prediction as applied to an unbalanced mixed linear model. *Biometrics*, **48**, 987–1003.

[184] Harville, D.A., Jeske, D.R. (1992) Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, **87**, 724–731.

[185] Harville, D.A., Zimmermann, A.G. (1996) The posterior distribution of the fixed and random effects in a mixed-effects linear model. *Journal of Statistical Computing and Simulation*, **54**, 211–229.

[186] Hauck, W.W., Anderson S. (1984) A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, **12**, 83–91.

[187] Hauck, W.W., Anderson S. (1986) A proposal for interpreting and reporting negative studies. *Statistics in Medicine*, **5**, 203–209.

[188] Hauck, W.W., Anderson S. (1992) Types of bioequivalence and related statistical considerations. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, **30**, 181–187.

[189] Hauck, W.W., Anderson S. (1994) Measuring switchability and presciptability: when is average bioequivalence sufficient. *Journal of Pharmacokinetics and Biopharmaceutics*, **22**, 551–564.

[190] Hauck, W.W., Anderson, S. (1999) Some issues in the design and analysis of equivalence trials. *Drug Information Journal*, **33**, 109–118.

[191] Hauck, W.W., Hyslop, T., Anderson, S., Bois, F.Y., Tozer, T.N. (1995) Statistical and regulatory considerations for multiple measures in bioequivalence testing. *Clinical Research and Regulatory Affairs*, **12**, 249–265.

[192] Hauck, W.W, Chen, ML., Hyslop, T., Patnaik, P., Shuirmann, D., Williams, R., for the FDA Individual Bioequivalence Working Group. (1996) Mean differences versus variability reduction: tradeoff in aggregate measures for individual bioequivalence. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, **34**, 535–541.

[193] Hauck, W.W., Hauschke, D., Diletti, E., Bois, F.Y., Steinijans, V.W., Anderson, S. (1997a) Choice of Student's t or Wilcoxon based confidence intervals for assessment of bioequivalence. *Journal of Biopharmaceutical Statistics*, **7**, 179–189.

[194] Hauck, W.W, Bois, F.Y., Hyslop, T., Gee, L., Anderson, S. (1997b) A parametric approach to population bioequivalence. *Statistics in Medicine*, **16**, 441–454.

[195] Hauck, W.W., Preston, P.E., Bois, F.Y. (1997c) A group sequential approach to cross-over trials for average bioequivalence. *Journal of Biopharmaceutical Statistics*, **7**, 87–96.

[196] Hauck, W.W., Hyslop, T., Chen, M.L., Patnaik, R., Williams, R.L., and the FDA Population/Individual Bioequivalence Working Group. (2000) Subject-by-formulation interaction in bioequivalence: Conceptual and statistical issues. *Pharmaceutical Research*, **17**, 375–380.

[197] Hauck, W.W., Parekh, A., Lesko, L., Chen, M-L., Williams, R. (2001) Limits of 80%-125% for AUC and 70%-143% for Cmax - What is the impact on bioequivalence studies? *International Journal of Clinical Pharmacology and Therapeutics*, **39**, 350–355.

[198] Hauschke, D., Steinijans, V.W., Diletti, E. (1990) A distribution free procedure for the statistical analysis of bioequivalence studies. *International Journal of Clinical Pharmacology, Therapy, and Toxicology*, **28**, 72–78.

[199] Hauschke, D., Hothorn, L. (1998) Safety assessment in toxicological studies: Proof of safety versus proof of hazard. *In Design and Analysis of Animal Studies in Pharmaceutical Development,* Chow, S., and Liu, J., eds., 197–226, Marcel Dekker, New York.

[200] Hauschke, D., Steinijans, V.W. (2000) The US draft guidance regarding population and individual bioequivalence approaches: Comments by a research-based pharmaceutical company. *Statistics in Medicine*, **19**, 2769–2774.

[201] Hauschke, D. (2002) Letter to the Editor and Correspondence on 'A Note on Sample Size Calculation in Bioequivalence Trials' by Chow and Wang (2001;28:155-169). *Journal of Pharmacokinetics and Pharmacodynamics*, **29**, 89–102.

[202] Haynes, J.D. (1981) Statistical simulation study of new proposed uniformity requirement for bioequivalency studies. *Journal of Pharmaceutical Sciences*, **70**, 673–675.

[203] Hellriegel, E.T., Bjornsson, T.D., Hauck, W.W. (1996) Interpatient variability in bioavailability is related to the extent of absorption: implications for bioavailability and bioequivalence studies. *Clinical Pharmacology and Therapeutics*, **60**, 601–607.

[204] Herchuelz, A. (1996) Bioequivalence assessment and the conduct of bioequivalence trials: a European point of view. *European Journal of Drug Metabolism and Pharmacokinetics*, **21**, 149–152.

[205] Hills, M., Armitage, P. (1979) The two period cross-over trial. *British Journal of Clinical Pharmacology*, **8**, 7–20.

[206] Hinkelmann, K., Kempthorne, O. (1994) *Design and Analysis of Experiments, Volume I: Introduction to Experimental Design*, John Wiley and Sons, New York.

[207] Hodges, J., Lehman, E. (1963) Estimates of location based on rank tests. *Annals of Mathematical Statistics*, **34**, 598–611.

[208] Hoenig, J., Heisey, D. (2001) The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, **55**, 19–24.

[209] Holder, D., Hsuan, F. (1993a) Moment based criteria for determining bioequivalence. *Biometrika*, **80**, 835–846.

[210] Holder, D., Hsuan, F. (1993b) A moment based criteria for determining individual bioequivalence. *Drug Information Journal*, **29**, 965–979.

[211] Holford, N.H.G., Hale, M., Ko, H.C., Steimer, J.-L., Sheiner, L.B., Peck, C.C. (1999) Simulation in drug development: good practices. *Center for Drug Development Science*, www.dml.georgetown.edu/cdds

[212] Hollander, M., Wolfe, D. (1999) *Nonparametric Statistical Methods.* John Wiley and Sons, New York.

[213] Honig, P., Woosley, R., Zamani, K., Conner, D., Cantilena, L. (1992) Changes in the pharmacokinetics and electrocardiographic pharmacodynamics of terfenadine with concomitant administration of erythromycin. *Clinical Pharmacology and Therapeutics*, **52**, 231–238.

[214] Howe, W.G. (1974) Approximate confidence limits on the mean of X+Y where X and Y are two tabled independent random variables. *Journal of the American Statistical Association*, **69**, 789–794.

[215] Hsu, J.C., Hwang, J.-T.G., Liu, H-K., Ruberg, S.J. (1994) Confidence limits associated with tests for bioequivalence. *Biometrika*, **81**, 103–114.

[216] Huitson, A. (1955) A method for assigning confidence limits to linear combinations of variances. *Biometrika*, **42**, 471–479.

[217] Hunt, M.I. (2000) Prescription Drugs and Intellectual Property Protection. *National Institute for Health Care Management: Foundation Issue Brief.*

[218] Huque, M.F., Dubey, S., Fredd, S. (1989) Establishing therapeutic equivalence with clinical endpoints. *Proceedings of the American Statistical Association, Biopharmaceutics Chapter*, 46–52.

[219] Huynh, H., Feldt, L.S. (1970) Conditions under which mean square ratios in repeated measures designs have exact F-distributions. *Journal of the American Statistical Association*, **65**, 1582–1589.

[220] Hwang, J.S. (1996) Numerical solutions for a sequential approach to bioequivalence. *Statistica Sinica*, **6**, 663–673.

[221] Hwang, J.T.G., Wang, W. (1997) The validity of the test of individual equivalence ratios. *Biometrika*, **84**, 893–900.

[222] Hwang, S., Huber, P.B., Hesney, M., Kwan, K.C. (1978) Bioequivalence and interchangeabiliy. *Journal of Pharmaceutical Sciences*, **67**, IV.

[223] Hyslop, T., Hsuan, F., Holder, D.J. (2000) A small sample confidence interval approach to assess individual bioequivalence. *Statistics in Medicine*, **19**, 2885–2897.

[224] Hyslop, T., Inglewicz, B. (2001) Alternative cross-over designs for individual bioequivalence. *Proceedings of the Annual Meeting of the American Statisticial Association.*

[225] International Conference on Harmonization (1994) Dose Response Information to Support Drug Registration. http://www.fda.gov/cder/guidance/

[226] International Conference on Harmonization (1995) Text on Validation of Analytical Procedures. http://www.fda.gov/cder/guidance/

[227] International Conference on Harmonization (1998) Guidance on Ethnic Factors in the Acceptability of Foreign Clinical Data. http://www.fda.gov/cder/guidance/

[228] International Conference on Harmonization (1998) Statistical Principles for Clinical Trials. http://www.fda.gov/cder/guidance/

[229] International Conference on Harmonization Consensus Guideline, Step 4 (2005) The Clinical Evaluation of QT/QTc Interval Prolongation and Proarrythmic Potential for Non-Antiarrythmic Drugs. http://www.fda.gov/cder/guidance/

[230] Jackson, A. (2000) The role of metabolites in bioequivalency assessment. III. Highly variable drugs with linear kinetics and first-pass effect. *Pharmaceutical Research*, **17**, 1432–1436.

[231] Jackson, A., Robbie, G., Marroum, P. (2004) Metabolites and bioequivalence - past and present. *Clinical Pharmacokinetics*, **43**, 655–672.

[232] Japan MHW (1997) Guideline for Bioequivalence Studies of Generic Products.

[233] Jennison, C., Turnbull, B.W. (2000) *Group Sequential Methods with Applications to Clinical Trials.* Chapman and Hall, New York.

[234] Jennrich, R.I., Schluchter, M.D. (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**, 805–820.

[235] Johnson, N.L., Kotz, S., Balakrishnan, N. (1994) *Continuous Univariate Distributions, Vols. 1 and 2.* John Wiley and Sons, New York.

[236] Johnston, A., Belitsky, P., Frei, U., Horvath, J., Hoyer, P., Helderman, J., Oellerich, M., Pollard, S., Riad, H., Rigotti, P., Keown, P., Nashan, B. (2004) Potential clinical implications of substitution of generic formulations for cyclosporine microemulsion (Neoral) in transplant recipients. *European Journal of Clinical Pharmacology*, **60**, 389–395.

[237] Jones, B., Kenward M.G. (2003) *Design and Analysis of Cross-over Trials, 2ed.* Chapman and Hall, London.

[238] Jones, B., Jarvis, P., Lewis, J.A., Ebbutt, A.F. (1996) Trials to assess equivalence: The importance of rigourous methods. *British Medical Journal*, **313**, 36–39.

[239] Jones, B., Lane, P.W. (2004). Procedure XOEFFICIENCY (Calculates efficiency of effects in cross-over designs). GenStat Release 7.1 Reference Manual, Part 3: Procedure Library PL15, VSN International, Oxford.

[240] Jones, B., Deppe, C. (2000) Recent developments in the design of cross-over trials: A brief review and bibliography. *Proceedings of the Conference on Recent Developments in the Design of Experiments and Related Topics.*

[241] Julious, S., Debarnot, C. A.-M. (2000) Why are pharmacokinetic data summarized by arithmetic means? *Journal of Biopharmaceutical Statistics*, **10**, 55–71.

[242] Julious, S., Patterson, S. (2004) Sample sizes for estimation in clinical research. *Pharmaceutical Statistics*, **3**, 213–215.

[243] Kanfer, F.-H.J., Geertsema, J.C., Steyn, H.S. (1988) A Two-Stage Procedure for Testing Bioequivalence. *Proceedings of the American Statistical Association*, 148–152.

[244] Kawai, N., Andoh, M., Uwoi, T., Goto, M. (2000) Statistical approaches to accepting foreign clinical data. *Drug Information Journal*, **34**, 1265–1272.

[245] Keiser, M., Hauschke, D. (2005) Assessment of clinical relevance by considering point estimates and associated confidence intervals. *Pharmaceutical Statistics*, **4**, 101–107.

[246] Kenward, M., Roger, J. (1997) Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, **33**, 983–997.

[247] Kepner, J.L., Randles, R.H. (1982) Detecting unequal marginal scales in a bi-variate population. *Journal of the American Statistical Association*, **77**, 475–482.

[248] Kershner, R.P., Federer, W.T. (1981) Two treatment cross-over designs for estimating a variety of effects. *Journal of the American Statistical Association*, **76**, 612–619.

[249] Kimanani, E.K., Potvin, D. (1997) A parametric confidence interval for a moment based scaled criterion for individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, **25**, 595–614.

[250] Kimanani, E.K., Lavigne, J., Potvin, D. (2000a) Numerical methods for the evaluation of individual bioequivalence. *Statistics in Medicine*, **19**, 2775–2795.

[251] Kimanani, E., Stypinski, D., Curtis, G., Stiles, M., Heessels, P., Logan, S., Melson, K., St Germain, E., Boswell, G. (2000b) A contract research organization's response to the new FDA guidances for bioequivalence/bioavailability studies for orally administered drug products. *Journal of Clinical Pharmacology*, **40**, 1102–1108.

[252] Kirkwood, T-B.L., Westlake, T.J. (1981) Letter to the editor and response. *Biometrics*, **37**, 589–593.

[253] Koch, G. (1972) The use of non-parametric methods in the statistical analysis of the two-period changeover design. *Biometrics*, **28**, 577–584.

[254] Kong, L., Koch, G., Liu, T., Wang, H. (2004) Performance of some multiple testing procedures to compare three doses of a test drug and placebo. *Pharmaceutical Statistics*, **4**, 25–35.

[255] Kullback, S. (1968) *Information Theory and Statistics.* Dover Publications, New York.

[256] Lacey, L.F., Keene, O.N., Bye, A. (1995) Glaxo's expreience of different absorption rate metrics of immediate release and extended release dosage forms. *Drug Information Journal*, **29**, 821–840.

[257] Lacey, L.F., Keene, O.N., Pritchard, J.F., Bye, A. (1997) Common noncompartmental pharmacokinetic variables: are they normally or log-normally distributed? *Journal of Biopharmaceutical Statistics*, **7**, 171–178.

[258] Lachenbruch, P. (2003) Proper metrics for clinical trials: transformations and other procedures to remove non-normality effects. *Statistics in Medicine*, **22**, 3823–3842.

[259] Laird, N.M., Ware, J.H. (1982) Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

[260] Lee, Y., Shao, J., Chow, S-C., Wang, H. (2002) Tests for intersubject and total variabilities under cross-over designs. *Journal of Biopharmaceutical Statistics*, **12(4)**, 503–534.

[261] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

[262] Lesko, L., Rowland, M., Peck, C., Blaschke, T. (2000) Optimizing the Science of Drug Development: opportunities for better candidate selection and accelerated evaluation in humans. *European Journal of Pharmaceutical Sciences*, **10**, iv–xiv.

[263] Lesko, L., Atkinson, A. (2001) Use of biomarkers and surrogate markers in drug development. *Annu. Rev. Pharmacol. Toxicol.*, **41**, 347–66.

[264] Levy, G. (1995) The clay feet of bioequivalence testing. *Journal of Pharmacy and Pharmacology*, **47**, 975–977.

[265] Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D. (1996) SAS System for Mixed Models. SAS Institute, Cary, North Carolina.

[266] Lin, L. (1989) A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–268.

[267] Lin, L. (1992) Assay validation using the concordance correlation coefficient. *Biometrics*, **48**, 599–604.

[268] Lin, L. (2000) Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine*, **19**, 255–270.

[269] Lindley, D.V. (1971) *Making Decisions, 2ed.* John Wiley and Sons, New York.

[270] Lindley, D.V, Singpurwalla, N.D. (1991) On the evidence needed to reach agreed action between adversaries, with application to acceptance sampling. *Journal of the American Statistical Association*, **86**, 933–937.

[271] Lindley, D.V. (1997) The choice of sample size (with discussion). *The Statistician*, **46**, 129–166.

[272] Lindley, D.V. (1998) Decision analysis and bioequivalence trials. *Statistical Science*, **13**, 136–141.

[273] Lindsey, J.K. (1996) *Parametric Statistical Inference.* Clarendon Press, Oxford.

[274] Lindsey, J.K. (2001) *Nonlinear Models in Medical Statistics.* Oxford University Press, Oxford.

[275] Lindsey, J.K., Jones, B. (1997) Treatment-patient interactions for diagnostics of cross-over trials. *Statistics in Medicine*, **16**, 1955–1964.

[276] Lindsey, J.K., Wang, J., Byrom, W.D., Jones, B. (1999) Modeling the covariance structure in pharmacokinetic cross-over trials. *Journal of Biopharmaceutical Statistics*, **9**, 439–450.

[277] Lindstrom, M.J., Bates, D.M. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models and repeated-measures data. *Journal of the American Statisticial Association*, **83**, 1014–1022.

[278] Liu, J.-P., Chow, S-C. (2002a) Bridging studies in clinical development. *Journal of Biopharmaceutical Statistics*, **12**, 359–367.

[279] Liu, J.-P., Hsiao, C-F., Hsueh, H. (2002b) Bayesian approach to evaluation of bridging studies. *Journal of Biopharmaceutical Statistics*, **12**, 401–408.

[280] Liu, J-P., Hsueh, H., Hsiao, C-F. (2004) A Bayesian noninferiority approach to evaluation of bridging studies. *Journal of Biopharmaceutical Statistics*, **14**, 291–300.

[281] Locke, C.S. (1984) An exact confidence interval for untransformed data for the ratio of two formulation means. *Journal of Pharmacokinetics and Biopharmaceutics*, **12**, 649–655.

[282] Longford, N.T. (1999) Selection bias and treatment heterogeneity in clinical trials. *Statistics in Medicine*, **18**, 1467–1474.

[283] Longford, N.T. (2000) An alternative definition of individual bioequivalence. *Statistica Neerlandica*, **14**, 14–36.

[284] Lu, T.C., Graybill, F.A., Burdick, R.K. (1988) Confidence intervals on a difference of expected mean squares. *Journal of Statistical Planning and Inference*, **18**, 35–43.

[285] Lund, R.E. (1975) Tables for an approximate test for outliers in linear models. *Technometrics*, **17**, 473–476.

[286] Machado, S., Miller, R., Hu, C. (1999) A regulatory perspective on pharmacokinetic and pharmacodynamic modelling. *Statistical Methods in Medical Research*, **8**, 217–245.

[287] Mallet, A. (1986) A maximum likelihood estimation method for random coefficient regression models. *Biometrika*, **73**, 645–656.

[288] Mancinelli, L., Frassetto, L., Floren, L., Dressler, D., Carrier, S., Bekersky, I., Benet, L., Christians, U. (2001) The pharmacokinetics and metabolic disposition of Tacrolimus: A comparison across ethnic groups. *Clinical Pharmacology and Therapeutics*, **69**, 24–31.

[289] Mandallaz, D., Mau, J. (1981) Comparison of different methods for decison making in bioequivalence assessment. *Biometrics*, **37**, 213–222.

[290] Marcus, R., Peritz, E., Gabriel, K.R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–60.

[291] Marston, S.A., Polli, J.E. (1997) Evaluation of direct curve comparison metrics applied to pharmacokinetic profiles and relative bioavailability and bioequivalence. *Pharmaceutical Research*, **14**, 1363–1369.

[292] Marzo, A., Monti, N.C., Tettamanti, R.A., Crivelli, F., Bo, L.D., Mazzucchelli, P., Meoli, A., Pezzuto, D., Corsico, A. (2000) Bioequivalence of inhaled Formoterol Fumurate assessed from pharmacodynamic, safety, and urinary pharmacokinetic data. *Arzneim-Forsch./Drug Res.*, **50**, 559–563.

[293] Matthews, J. (1994) Modelling and optimality in the design of cross-over studies for medical applications. *Journal of Statistical Planning and Inference*, **42**, 89–108.

[294] Mauger, D.T., Chinchilli, V.M. (2000) An alternative index for assessing profile similarity in bioequivalence trials. *Statistics in Medicine*, **19**, 2855–2866.

[295] McCulloch, C.E. (1987) Tests for equality of variances with paired data. *Communications in Statistical Theory and Methods*, **16**, 1377–1391.

[296] Meibohm, B., Derendorf, H. (2002) Pharmacokinetic-pharmacodynamic studies in drug product development. *Journal of Pharmaceutical Sciences*, **91**, 18–31.

[297] Metzler, C.M. (1974) Bioavailability: A problem in equivalence. *Biometrics*, **30**, 309–317.

[298] Metzler, C.M., Huang, D.C. (1983) Statistical methods for bioavailability and bioequivalence. *Clinical Research Practices and Drug Regulatory Affairs*, **1**, 109–132.

[299] Meyer, M.C. (1995) Current scientific issues regarding bioavailability/bioequivalence studies: An academic view. *Drug Information Journal*, **29**, 805–812.

[300] Meyer, M.C., Straughn, A.B., Jarvi, E.J., Patrick, K.S., Pelsor, F.R., Williams, R.L., Patnaik, R., Chen, M.L., Shah, V.P. (2000) Bioequivalence of Methylphenidate immediate-release tablets using a replicated study design to characterize intra-subject variability. *Pharmaceutical Research*, **17**, 381–384.

[301] Meyer, M.C., Straughn, A.B., Mhatre, R.M., Shah, V.P., Chen, M.-L., Williams, R.L., Lesko, L.J. (2001) Variability in the bioavailability of Phenytoin capsules in males and females. *Pharmaceutical Research*, **18**, 394–397.

[302] Midha, K.K., Ormsby, E.D., Hubbard, J.W., McKay, G., Hawes, E.M., Gavalas, L., McGilveray, I.J. (1993) Logarithmic transformation in bioequivalence: Application with two formulations of Perphenazine. *Journal of Pharmaceutical Sciences*, **82**, 138–144.

[303] Midha, K.K., Rawson, M.J., Hubbard, J.W. (1997a) Individual and average bioequivalence of high variability drugs and drug products. *Journal of Pharmaceutical Sciences*, **86**, 1193–1197.

[304] Midha, K.K., Rawson, M.J., Hubbard, J.W. (1997b) Bioequivalence: switchability and scaling. *European Journal of Pharmaceutical Sciences*, **6**, 87–91.

[305] Millard, S.P., Krause, A. (2001) *Applied Statistics in the Pharmaceutical Industry.* Springer, New York.

[306] Miller, M. (2001) Gender-based differences in the toxicity of pharmaceuticals - The Food and Drug Administration's perspective. *International Journal of Toxicology*, **20**, 149–152.

[307] Milliken, G.A., Johnson, D.E. (1992) *Analysis of Messy Data, Vol 1: Designed Experiments.* Chapman and Hall, New York.

[308] Morgan, W.A. (1939) A test for the significance of the difference between two variances in a sample from a bi-variate normal population. *Biometrika*, **31**, 13–19.

[309] Morrison, D. (1990) *Multivariate Statistical Methods, 3ed.* McGraw Hill, New York.

[310] Moss, A. (1993) Measurement of the QT interval and the risk associated with QT interval prolongation. *American Journal of Cardiology*, **72**, 23B–25B.

[311] Muirhead, R.J. (1982) *Aspects of Multivariate Statistical Theory.* John Wiley and Sons, New York.

[312] Munk, A. (1993) An improvement on commonly used tests in bioequivalence assessment. *Biometrics*, **49**, 1225–1230.

[313] Munk, A., Pfluger, R. (1999) 1-α equivalence confidence rules for convex alternatives are $\alpha/2$ tests with applications to the multivariate assessment of bioequivalence. *Journal of the American Statistical Association*, **94**, 1311–1319.

[314] Myers, R.H. (1990) *Classical and Modern Regression with Applications, 2ed.* PWS-Kent, Boston.

[315] Nagata, R., Fukase, H., Rafizadeh-Kabe, J. (2000) East-West development: Understanding the usability and acceptance of foreign data in Japan. *International Journal of Clinical Pharmacology and Therapeutics*, **38**, 87–92.

[316] Naito, C. (1998a) Relevance of ethnic factors in global development: Historic overview. *In Proceedings of the Fourth International Conference on Harmonisation*, D'Arcy, PF., Harron, D. eds., Greystone Books.

[317] Naito, C. (1998b) Ethnic factors in the acceptability of foreign clinical data. *Drug Information Journal*, **32**, 1283S–1292S.

[318] Ormsby, E.D. (1999) Invited Presentation: Individual Bioequivalence: A Canadian Perspective. *AAPS International Workshop on Individual Bioequivalence: Realities and Implementation.*

[319] O'Quigley, J., Baudoin, C. (1988) General approaches to the problem of bioequivalence. *The Statistician*, **37**, 51–58.

[320] O'Quigley, J., Pepe, M., Fisher, L. (1990) Continual reassessment method: A practical design for phase 1 studies in cancer. *Biometrics*, **46**, 33–48.

[321] Owen, D.B. (1965) A special case of a bi-variate non-central t-distribution. *Biometrika*, **52**, 437–446.

[322] Pabst, G., Jaeger, H. (1990) Review of methods and criteria for the evaluation of bioequivalence studies. *European Journal of Clinical Pharmacology*, **38**, 5–10.

[323] Patnaik, P.B. (1949) The non-central χ^2 and F-distributions and their applications. *Biometrika*, **36**, 202–232.

[324] Patnaik, R., Lesko, L.J., Chan, K., Williams, R.L. (1996) Bioequivalence assessment of generic drugs: An American point of view. *European Journal of Drug Metabolism and Pharmacokinetics*, **21**, 159–164.

[325] Patnaik, R.N., Lesko, L.J., Chen, ML., Williams, R.J., and the FDA Individual Bioequivalence Working Group. (1997) Individual bioequivalence–new concepts in the statistical assessment of bioequivalence metrics. *Clinical Pharmacokinetics*, **33**, 1–6.

[326] Patterson, H.D. (1950) The construction of balanced designs for experiments involving sequences of treatments. *Biometrika*, **39**, 32.

[327] Patterson, H.D., Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

[328] Patterson, S., Francis, S., Ireson, M., Webber, D., Whitehead, J. (1999) A novel Bayesian decision procedure for early-phase dose-finding studies. *Journal of Biopharmaceutical Statistics*, **9**, 583–598.

[329] Patterson, S. (2001a) A review of the development of biostatistical design and analysis techniques for assessing in vivo bioequivalence, Part 1. *Indian Journal of Pharmaceutical Sciences*, **63**, 81–100.

[330] Patterson, S. (2001b) A review of the development of biostatistical design and analysis techniques for assessing in vivo bioequivalence, Part 2. *Indian Journal of Pharmaceutical Sciences*, **63**, 169–186.

[331] Patterson, S., Zariffa, N., Howland, K., Montague, T. (2001c) Non-traditional study designs to demonstrate average bioequivalence for highly variable drug products. *European Journal of Clinical Pharmacology*, **57**, 663–670.

[332] Patterson, S., Jones, B. (2002a) Bioequivalence and the pharmaceutical industry. *Pharmaceutical Statistics*, **1**, 83–95.

[333] Patterson, S., Jones, B. (2002b) Statistical aspects of bioequivalence in the pharmaceutical industry. *Proceedings of the American Statistical Association Joint Statistical Meetings.*

[334] Patterson, S., Jones, B. (2002c) Clinical development planning and the use of pharmacokinetic data in ICH-E5 bridging assessments. *GSK BDS Technical Report 2002-04.*

[335] Patterson, S., Jones, B. (2004a) Simulation assessments of statistical aspects of bioequivalence in the pharmaceutical industry. *Pharmaceutical Statistics*, **3**, 13–23.

[336] Patterson, S., Agin, M., Anziano, R., Chuang-Stein, C., Dmitrienko, A., Ferber, G., Francom, S., Geraldes, M., Ghosh, K., Mills, T., Menton, R., Natarajan, J., Offen, W., Saoud, J., Smith, B., Suresh, R., Zariffa, N. (In Press) Investigating Drug Induced QT and QTc Prolongation in the Clinic: Statistical Design and Analysis Considerations. *Drug Information Journal.*

[337] Peace, K.E. (1986) Estimating the degree of equivalence and non-equivalence, an alternative to bioequivalence testing. *Proceedings of the American Statistical Association*, 63–69.

[338] Peace, K., ed. (1992) Biopharmaceutical Sequential Statistical Applications. Dekker, New York.

[339] Peace, K.E. (1993) Design and analysis considerations for safety data, particularly adverse events. *In Drug Safety Assessment in Clinical Trials*, Gibert, G.S. ed., 305–316. Marcel Dekker, New York.

[340] Peck, C., Desjardins, R. (1996) Simulation of clinical trials: Encouragements and cautions. *Applied Clinical Trials*, 30–32.

[341] Phillips, K.F. (1990) Power of the two one-sided testing procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, **18**, 137–144.

[342] Pitman, E.C.G. (1939) A note on normal correlation. *Biometrika*, **31**, 9–12.

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

[344] Pound, N.J. (1999) Bioavailability and Bioequivalence: Update on Guidances from TPP. *AAPS International Workshop on Individual Bioequivalence: Realities and Implementation.*

[345] Pradhan, R.S. (1997) Role of interoccasion variation in estimation of bioequivalence: A Bayesian approach. *Clinical Pharmacology and Therapeutics*, **61**, 186.

[346] Pratt, C., Hertz, R., Elis, B., Crowell, S., Louv, W., Moye, L. (1994) Risk of developing life-threatening ventricular arrhythmia associated with terfenadine in comparison with other over the counter antihistamines. *American Journal of Cardiology*, **73**, 346–352.

[347] Pratt, C., Ruberg, S., Morganroth, J., McNutt, B., Woodward, J., Harris, S., Ruskin, J., Moye, L. (1996) Dose-response relation between terfenadine (Seldane) and the QTc interval on the scalar electrocardiogram: Distinguishing drug effect from spontaneous variability. *American Heart Journal*, **131**, 472–480.

[348] Prentice, R. (1989) Surrogate endpoints in clinical trials: Definition and operational criteria. *Statistics in Medicine*, **8**, 431–440.

[349] Priori, S., Schwartz, P., Napolitano, C., Bloise, R., Ronchetti, E., Grillo, M., Vicentini, A., Spazzolini, C., Nastoli, J., Bottelli, G., Folli, R., Capelletti, D. (2003) Risk stratification in the long-QT syndrome. *New England Journal of Medicine*, **348**, 1866–1874.

[350] Quiroz, J., Ting, N., Wei, G., Burdick, R. (2002) Alternative confidence intervals for the assessment of bioequivalence in four-period cross-over designs. *Statistics in Medicine*, **21**, 1825–1847.

[351] Racine-Poon, A., Grieve, A.P., Fluhler, H., Smith, A. F-M. (1986) Bayesian methods in practice, experiences in the pharmaceutical industry (with discussion). *Applied Statistics*, **35**, 93–150.

[352] Racine-Poon, A., Grieve, A.P., Fluhler, H., Smith, A. F-M. (1987) A two-stage procedure for bioequivalence studies. *Biometrics*, **43**, 847–856.

[353] Racoosin, J. (2003) The clinical evaluation of QT interval prolongation and proarrythmic potential for non-antiarrythmic Drugs. *Presentation at Drug Information Agency/FDA Workshop*, www.diahome.org

[354] Rao, C.R. (1973) *Linear Statistical Inference and Its Applications, 2ed.* John Wiley and Sons, New York.

[355] Reigner, B., Williams, P., Patel, I., Steimer, J., Peck, C., Brummelen, P. (1997) An evaluation of the integration of pharmacokinetic and pharmacodynmaic principles in clinical drug development. *Clinical Pharmacokinetics*, **33**, 142–152.

[356] Reigner, B., Blesch, K. (2001) Estimating the starting dose for entry into humans: principles and practice. *European Journal of Clinical Pharmacology*, **57**, 835–845.

[357] Reiser, B., Guttman, I. (1986) Statistical inference for $\Pr(Y < X)$. *Technometrics*, **28**, 253–257.

[358] Rescigno, A. (1992) Bioequivalence. *Pharmaceutical Research*, **9**, 925–928.

[359] Rescigno, A., Powers, J.D. (1998) AUC and Cmax are not sufficient to prove bioequivalence. *Pharmacological Research*, **37**, 93–95.

[360] Rheinstein, P.H. (1990) Therapeutic inequivalence. *Drug Safety 5*, **Suppl 1**, 114–119.

[361] Rhodes, C.T. (1997) Acceptance limits for bioequivalence studies. *Clinical Research and Regulatory Affairs*, **14**, 127–137.

[362] Ring, A., Tothfalusi, L., Endrenyi, L., Weiss, M. (2000) Sensitivity of empirical metrics of rate of absorption in bioequivalence studies. *Pharmaceutical Research*, **17**, 583–588.

[363] Rocke, D.M. (1984) On testing for bioequivalence. *Biometrics*, **40**, 225–230.

[364] Rodda, B.E., Davis R.L. (1980) Determining the probability of an important difference in bioavailability. *Clinical Pharmacology and Therapeutics*, **28**, 247–252.

[365] Roe, D., Vonesh, E., Wolfinger, R., Mensil, F.,Mallet, A. (1997) Comparison of population pharmacokinetic modelling methods using simulated data: Results from the Population Modelling Workgroup. *Statistics in Medicine*, **16**, 1241–1262.

[366] Rowland, M., Tozer, T.N. (1980) *Clinical Pharmacokinetics: Concepts and Applications.* Lea and Febidger, Philadelphia.

[367] Sarkar, S., Watts, S., Ohashi, O., Carroll, R., Uesaka, H., Mason, T., Rivera, C. (2002) Bridging data between two ethnic populations. A new application of matched case-control methodology. *Drug Information Journal*, **36**, 349–356.

[368] *Statistical Analysis Software, SAS/STAT User's Guide, Version 8.* (1999). SAS Institute, Cary, NC.

[369] Satterthwaite, F. (1941) Synthesis of variance. *Psychometrika*, **6**, 309–316.

[370] Schall, R. (1995) Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics*, **51**, 615–626.

[371] Schall, R., Luus, H.G. (1993) On population and individual bioequivalence. *Statistics in Medicine*, **12**, 1109–1124.

[372] Schall, R., Williams, R.L. (1996) Towards a practical strategy for assessing individual bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, **24**, 133–149.

[373] Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6(2)**, 461–464.

[374] Schuirmann, D.J. (1981) On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics*, **37**, 617.

[375] Schuirmann, D.J. (1987) A comparison of the two one sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657–680.

[376] Schuirmann, D.J. (1990) Design of bioavailability and bioequivalence studies. *Drug Information Journal*, **24**, 315–323.

[377] Schulman, R.L. (1992) *Statistics in Plain English with Computer Applications*. Van Nostrand Reinhold, New York.

[378] Schumaker, R.C., Metzler, C.M. (1998) The phenytoin trial is a case study of individual bioequivalence. *Drug Information Journal*, **32**, 1063–1072.

[379] Searle, S.R. (1971) *Linear Models*. John Wiley and Sons, New York.

[380] Seber, G.A. (1977) *Linear Regression Analysis*. John Wiley and Sons, New York.

[381] Selwyn, M.R., Dempster, A.P., Hall, N.R. (1981) A Bayesian approach to bioequivalence for the 2 × 2 cross-over design. *Biometrics*, **37**, 11–21.

[382] Selwyn, M.R., Hall, N.R. (1984) On Bayesian methods for bioequivalence. *Biometrics*, **40**, 1103–1108.

[383] Senn, S. (1996) The AB/BA cross-over: How to perform the two-stage analysis if you can't be persuaded you shouldn't. *In Liber Amicorum Roel van Strik*, Hansen, B., and de Ridder, M. eds., 93–100. Erasmus University, Rotterdam.

[384] Senn, S. (1997) *Statistical Issues in Drug Development*. John Wiley and Sons, New York.

[385] Senn, S. (1998) In the blood: Proposed new requirements for registering generic drugs. *The Lancet*, **352**, 85–86.

[386] Senn, S. (2000) Decisions and Bioequivalence. *Conference Proceedings of Challenging Statistical Issues in Clinical Trials.*

[387] Senn, S. (2001) Statistical issues in bioequivalence. *Statistics in Medicine*, **20**, 2785–2799.

[388] Senn, S. (2002) *Cross-over Trials in Clinical Research, 2ed.* John Wiley and Sons, New York.

[389] Senn, S., Lee, S. (2004a) The analysis of the AB/BA cross-over trial in the medical literature. *Pharmaceutical Statistics*, **3**, 123–131.

[390] Senn, S., D'Angelo, G., Potvin, D. (2004b) Carry-over in cross-over trials in bioequivalence: Theoretical concerns and empirical evidence. *Pharmaceutical Statistics*, **3**, 133–142.

[391] Serfling, R. (1980) *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, New York.

[392] Shaffer, J. (1986) Modified Sequentially Rejective Multiple Test Procedures. *Journal of the American Statistical Association*, **81**, 826–831.

[393] Shah, V.P., Yacobi, A., Barr, W., Benet, L.Z., Breimer, D., Dobrinska, M.R., Endrenyi, L., Fairweather, W., Gillespie, W., Gonzalez, M.A., Hooper, J., Jackson, A., Lesko, L.J., Midha, K.K., Noonan, P.K., Patnaik, R., Williams, R.L. (1996) Workshop report: Evaluation of orally administered highly variable drugs and drug formulations. *Pharmaceutical Research*, **13**, 1590–1594.

[394] Shao, J., Tu, D. (1996) *The Jackknife and Bootstrap*. Springer, New York.

[395] Shao, J., Chow, S.-C., Wang, B. (2000a) The bootstrap procedure in individual bioequivalence. *Statistics in Medicine*, **19**, 2741–2754.

[396] Shao, J., Kubler, J., Pigeot, I. (2000b) Consistency of the bootstrap procedure in individual bioequivalence. *Biometrika*, **87**, 573–585.

[397] Shapiro, S., Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

[398] Sharma, A., Pilote, S., Belanger, P., Arsenault, M., Hamelin, B. (2004). A convenient five-drug cocktail for the assessment of major drug metabolizing enzymes: A pilot study, *British Journal of Clinical Pharmacology*, **58**, 288–297.

[399] Sheiner L.B., Beal, S.L., Sambol, N.C. (1989) Study designs for dose-ranging. *Clinical Pharmacology and Therapeutics*, **46**, 63–77.

[400] Sheiner, L.B., Hashimoto, Y., Beal, S. (1991) A simulation study comparing designs for dose ranging. *Statistics in Medicine*, **10**, 303–321.

[401] Sheiner, L.B. (1992) Bioequivalence revisited. *Statistics in Medicine*, **11**, 1777–1788.

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

[403] Sheiner, L.B., Steimer, J-L. (2000) Pharmacokinetic-Pharmacodynamic Modeling in Drug Development. *Annu. Rev. Pharmacol. Toxicol.*, **40**, 67–95.

[404] Shih, W. (2001) Clinical trials for drug registrations in Asian-Pacific countries: Proposal for a new paradigm from a statistical perspective. *Controlled Clinical Trials*, **22**, 357–366.

[405] Singh, G.J.-P., Adams, W.P., Lesko, L.J., Shah, V.P., Molzon, J.A., Williams, R.L., Pershing, L.K. (1999) Development of in vivo bioequivalence methodology for dermatologic corticosteroids based on pharmacodynamic modeling. *Clinical Pharmacology and Therapeutics*, **66**, 346–357.

[406] Smith, B., Vandenhende, F., DeSante, K., Farid, N., Welch, P., Callaghan, J., Forgue, S. (2000) Confidence interval criteria for assessment of dose proportionality. *Pharmaceutical Research*, **17**, 1278–1283.

[407] Smith, M. (2003) Software for non-linear mixed effects modelling: a review of several packages. *Pharmaceutical Statistics*, **2**, 69–75.

[408] Smith, B. (2004) Assessment of Dose Proportionality. *In Pharmacokinetics in Drug Development: Clinical Study Design and Analysis, Volume 1*, Bonate, P., Howard, D., eds., 363–382. AAPS Press, USA.

[409] Snikeris, F., Tingey, H.B. (1994) A two step method for assessing bioequivalence. *Drug Information Journal*, **28**, 709–722.

[410] Spino, M., Tsang, Y.C., Pop, R. (2000) Dissolution and in vivo evidence of differences in reference products: Impact on development of generic drugs. *European Journal of Drug Metabolism and Pharmacokinetics*, **25**, 18–24.

[411] Sprecher, D., Watkins, T., Behar, S., Brown, W., Rubins, H., Schaefer, E. (2003) Importance of high density lipoprotein cholesterol and triglyceride levels in coronary heart disease. *The American Journal of Cardiology*, **91**, 575–580.

[412] Srinivasan, R, Langenberg, P. (1986) A two-stage procedure with controlled error probabilities for testing equivalence. *Biometrical Journal*, **7**, 825–833.

[413] Stedman's Medical Dictionary, 25th ed. (1990) William and Wilkins, Baltimore.

[414] Steinijans, V.W., Diletti, E. (1983) Statistical analysis of bioavailability studies: Parametric and nonparametric confidence interval. *European Journal of Clinical Pharmacology*, **24**, 127–136.

[415] Steinijans, V.W., Hauschke, V. (1990) Update on the statistical analysis of bioequivalence studies. *International Journal of Clinicial Pharmacology, Therapy, and Toxicology*, **28**, 105–110.

[416] Steinijans, V.W., Hartmann, M., Huber, R., Radtke, H.W. (1991) Lack of pharmacokinetic interaction as an equivalence problem. *International Journal of Clinicial Pharmacology, Therapy, and Toxicology*, **29**, 323–328.

[417] Steinijans, V.W., Sauter, R., Hauschke, D., Elze, M. (1995) Metrics to characterize concentration-time profiles in single and multiple-dose bioequivalence studies. *Drug Information Journal*, **29**, 981–987.

[418] Steinijans, V.W., Hauschke, D., Schall, R. (1995) International harmonization of regulatory requirements for average bioequivalence and current issues in individual bioequivalence. *Drug Information Journal*, **29**, 1055–1062.

[419] Steinjans, V.W., Diletti, E. (1997) Individual bioequivalence: A European perspective. *Journal of Biopharmaceutical Statistics*, **7**, 31–34.

[420] Stokes, M., Davis, C., Koch, G. (2002) *Categorical Data Analysis Using the SAS System*. SAS Institute, Cary, NC.

[421] Strom, B.L. (1987) Generic drug substitution revisited. *NEJM*, **316**, 1456–1462.

[422] Suganami, H. (2004) Points to notice and proposal drug induced QT interval prolongation - Is your correction good enough? *Drug Information Association Meeting,* www.diahome.org

[423] Swallow, W., Monahan, J. (1984) Monte carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components. *Technometrics*, **26**, 47–57.

[424] Tanida, N. (2002) Ethical considerations for clinical trials in Asia. *Drug Information Journal*, **36**, 41–9.

[425] Temple, R. (1999) Are surrogate markers adequate to assess cardiovascular disease drugs? *Journal of the American Medical Association*, **282**, 790–795.

[426] Temple, R. (2003) Overview of the concept paper, history of the QT/TdP concern; Regulatory implications of QT prolongation. *Presentations at Drug Information Agency/FDA Workshop*, www.diahome.org

[427] Thall, P., Cook, J. (2004) Dose-finding based on efficacy-toxicity trade-offs. *Biometrics*, **60**, 684–693.

[428] Ting, N., Burdick, R.K., Graybill, F.A., Jeyaratnam, S., Lu, T-F.C. (1990) Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computing and Simulation*, **35**, 135–143.

[429] Tothfalusi, L., Endrenyi, L., Midha, K., Rawson, M., Hubbard, J. (2001) Evaluation of the bioequivalence of highly-variable drugs and drug products. *Pharmaceutical Research*, **18**, 728–733.

[430] Tothfalusi, L., Endrenyi, L. (2003) Limits for the scaled average bioequivalence of highly variable drugs and drug products. *Pharmaceutical Research*, **20**, 382–389.

[431] Tozer, T.N., Bois, F.Y., Hauck, W.W., Chen, M.L., Williams, R.L. (1996) Absorption rate versus exposure: Which is more useful for bioequivalence testing? *Pharmaceutical Research*, **13**, 453–456.

[432] Tozer, T.N., Hauck, W.W. (1997) Cmax/AUC, a commentary. *Pharmaceutical Research*, **14**, 967–968.

[433] Tsang, Y.C., Opo, R., Gordon, P., Hems, J., Spino, M. (1996) High variability in drug pharmacokinetics complicates determination of bioequivalence: Experience with Verapamil. *Pharmaceutical Research*, **13**, 846–850.

[434] Tudor, G., Koch, G. (1994) Review of nonparametric methods for the analysis of crossover studies. *Statistical Methods in Medical Research*, **3**, 345–381.

[435] Vonesh, E.F., Chinchilli, V.M. (1997) *Linear and Nonlinear Models for the Analysis of Repeated Measurements.* Marcel Dekker, New York.

[436] Vuorinen, J. (1997) A practical approach for the assessment of bioequivalence under selected higher order cross-over designs. *Statistics in Medicine*, **16**, 2229–2243.

[437] Vuorinen, J., Turunen, J. (1996) A three step procedure for assessing bioequivalence in the general mixed linear framework. *Statistics in Medicine*, **15**, 2635–2655.

[438] Wakefield, J.C., Smith, A.F.-M., Racine-Poon, A., Gelfand, A.E. (1994) Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, **43**, 201–221.

[439] Walpole, R.E., Myers, R.H., Myers, S.L. (1998) *Probability and Statistics for Engineers and Scientists, 6ed.* Prentice Hall, New Jersey.

[440] Wang, C.M. (1990) On the lower bound of confidence coefficients for a confidence interval on variance components. *Biometrics*, **46**, 187–192.

[441] Wang, W. (1997) Optimal unbiased tests for equivalence in intrasubject variability. *Journal of the American Statistical Association*, **88**, 939–946.

[442] Wang, W. (1999) On testing for individual bioequivalence. *Journal of the American Statistical Association*, **94**, 880–887.

[443] Wang, Y. (1999) Use of jacknife influence profiles in bioequivalence evaluations. *Pharmaceutical Science and Technology Today*, **2**, 152–159.

[444] Welleck, S. (1993) Basing the analysis of comparative bioavailability trials on an individualized statistical definition of equivalence. *Biometrical Journal*, **35**, 47–55.

[445] Welleck, S. (1997) A comment on so called individual criteria of bioequivalence. *Journal of Biopharmaceutical Statistics*, **7**, 17–21.

[446] Welleck, S. (2000) On a reasonable disaggregate criterion of population bioequivalence admitting of resampling-free testing procedures. *Statistics in Medicine*, **19**, 2755–2767.

[447] Welleck, S. (2003) *Testing Statistical Hypotheses of Equivalence.* Chapman and Hall, London.

[448] Welty, T.E., Pickering, P.R., Hale, B.C., Arazi, R. (1992) Loss of seizure control associated with generic substitution of carbamazepine. *Annals of Pharmacotherapy*, **26**, 775–777.

[449] Westfall, P. (1997) Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, **92**, 299–306.

[450] Westfall, P., Tobias, R., Rom, D., Wolfinger, R., Hochberg, Y. (1999) *Multiple Comparisons and Multiple Tests using SAS.* SAS Institute, Cary, NC.

[451] Westlake, W.J. (1972) Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences*, **61**, 1340–1341.

[452] Westlake, W.J. (1979) Statistical aspects of comparative bioavilability trials. *Biometrics*, **35**, 273–280.

[453] Westlake, W.J. (1986) Bioavailability and bioequivalence of pharmaceutical formulations. *In Biopharmaceutical Statistics for Drug Development*, Peace, K. ed., 329–352. Marcel Dekker, New York.

[454] Wetherill, G., Glazebrook, K. (1986) *Sequential Methods in Statistics.* Chapman and Hall, New York.

[455] Whitehead, J. (1996) Sequential designs for equivalence studies. *Statistics in Medicine*, **15**, 2703–2715.

[456] Whitehead, J., Zhou, Y., Stallard, N., Todd, S., Whitehead, A. (2001) Learning from previous responses in phase 1 dose escalation studies. *British Journal of Clinical Pharmacology*, **52**, 1–7.

[457] Willavize, S., Morgenthein, E. (2005) Comparison of models for average bioequivalence in replicate designs. *Proceedings of the American Statistical Association.*

[458] Williams, R.L., Adams, W., Chen, M.-L., Hare, D., Hussain, A., Lesko, L., Patnaik, R., Shah, V., and FDA Biopharmaceutics Coordinating Committee (2000a) Where are we now and where do we go next in terms of the scientific basis for regulation on bioavailability and bioequivalence? *European Journal of Drug Metabolism and Pharmacokinetics*, **25**, 7–12.

[459] Williams, R.L., Patnaik, R.N., Chen, M.-L. (2000b) The basis for individual bioequivalence. *European Journal of Drug Metabolism and Pharmacokinetics*, **25**, 13–17.

[460] Wolfinger, R., Tobias, R., Sall, J. (1994) Computing gaussian likelihoods and their derivatives for general linear mixed models. *Siam J Sci Comput*, **15**, 1294–1310.

[461] Wolfinger, R.D., Kass, R.E. (2000) Nonconjugate Bayesian analysis of variance component models. *Biometrics*, **56**, 768–774.

[462] Wysowski, D., Corken, A., Gallo-Torres, H., Talarico, L., Rodriguez, E. (2001) Postmarketing reports of QT prolongation and ventricular arrhythmia in association with Cisapride and FDA regulatory actions. *American Journal of Gastroenterology*, **96**, 1698–1703.

[463] Yacobi, A., Masson, E., Moros, D., Ganes, D., Lapointe, C., Abolfathi, Z., LeBel, M., Golander, Y., Doepner, D., Blumberg, T., Cohen, Y., Levitt, B. (2000) Who needs individual bioequivalence studies for narrow therapeutic index drugs? A case for warfarin. *Journal of Clinical Pharmacology*, **40**, 826–835.

[464] Yasuhara, H. (1994) Which is more important in pharmacokinetics: inter-ethnic variability or intra-ethnic variability. *In Proceedings of the Second International Conference on Harmonisation, 1993*, D'Arcy, PF., Harron, D., eds.

[465] Yamaoka, K., Nakagawa, T., Uno, T. (1978) Statistical moments in pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics*, **6**, 547–558.

[466] Yee, K.F. (1986) The calculation of probabilities in rejecting bioequivalence. *Biometrics*, **42**, 961–965.

[467] Yeh, K.C., Kwan, K.C. (1978) A comparison of numerical integrating algorithms of trapezoidal, Legrange, and spline approximation. *Journal of Pharmacokinetics and Biopharmaceutics*, **6**, 79–81.

[468] Yuh, L., Beal, S., Davidian, M., Harrison, F., Hester, A., Kowalski, K., Vonesh, E., Wolfinger, R. (1994) Population pharmacokinetic-pharmacodynamic methodology and applications: A bibliography. *Biometrics*, **50**, 566–575.

[469] Zariffa, N., M.-D., Patterson, S.D., Boyle, D., Hyneck, H. (1998) Case Studies, comparison of the current and proposed bioequivalence criteria. *FDA/AAPS Workshop on Bioavailability and Bioequivalence.*

[470] Zariffa, N., M.-D., Patterson, S.D. (1999) A case review on the topic of subject by formulation interaction. *AAPS International workshop on Individual Bioequivalence: Realities and Implementation.*

[471] Zariffa, N., M.-D., Patterson, S.D. (2000) Learnings and recommendations for population and individual bioequivalence based on the analysis of replicate study designs. *AAPS/FDA Workshop on Biopharmaceutics in the New Millennium: Regulatory Approaches to Bioavailability and Bioequivalence.*

[472] Zariffa, N., M.-D., Patterson, S.D., Boyle, D., Hyneck, H. (2000) Case studies, practical issues, and observations on population and individual bioequivalence. *Statistics in Medicine*, **19**, 2811–2820.

[473] Zariffa, N.M–D., Patterson, S.D. (2001) Population and individual bioequivalence: Lessons from real data and simulation studies. *Journal of Clinical Pharmacology*, **41**, 811–822.